EÖTVÖS LORÁND TUDOMÁNYEGYETEM
PEDAGÓGIAI ÉS PSZICHOLÓGIAI KAR

Takácsné Kárász Judit

Adaptív teljesítménymérési algoritmusok kidolgozása az Országos kompetenciamérés adatainak felhasználásával

DOI-azonosító: 10.15476/ELTE.2024.139

Neveléstudományi Doktori Iskola
A Doktori Iskola vezetője: Dr. habil. Zsolnai Anikó, egyetemi tanár

Oktatás-tanulás-egyenlőtlenségek program
Programvezető: Dr. habil. Lénárd Sándor, egyetemi docens

Témavezetők: Dr. habil. Nahalka István CSc, ny. egyetemi docens
Dr. habil. Széll Krisztián László, egyetemi docens

Budapest, 2024

Tartalom

Ábrajegyzék

Táblázatok jegyzéke

Köszönetnyilvánítás

Ez a disszertáció több, mint öt évnyi munka gyümölcse, mely sokak támogatásával jöhetett létre. Köszönöm a témavezetőimnek, Széll Krisztiánnak és Nahalka Istvánnak, akkor és úgy segítettek, amikor szükségem volt rá. Nagyszerűen kiegészítették egymást, mindkettőjüktől mást és mást tanultam. Hálás vagyok a segítségükért.

Köszönöm a korábbi programvezetőknek, Vámos Ágnesnek és Szivák Juditnak, hogy bizalmat szavaztak a szokatlan témának, segítettek, támogattak ezen a számomra idegen pályán. Sajnálom, hogy nem láthatják e munka végeredményét. Köszönöm a PPP kutatócsoportnak, hogy befogadott, és annak ellenére, hogy témáink sokszor igencsak távol álltak egymástól, támogató légkörben megosztották velem gondolataikat, észrevételeiket.

Köszönöm az Oktatási Hivatalnak, főleg Szepesi Ildikónak és Ostorics Lászlónak, hogy engedélyezték és támogatták a doktori tanulmányok megkezdését. Rugalmasságuk nélkül nem juthattam volna el idáig. Köszönöm az ÚNKP-nak, hogy három alkalommal is anyagi támogatást nyújtottak, ez végig segítette a kutatások megvalósulását. A leadási határidők inspirálólag hatottak, egyben az elvégzett munka látható mérföldköveivé váltak.

Köszönöm Nikinek, Orsinak és Fruzsinak, hogy doktoranduszokként társaim voltak az úton, megoszthattam és ti is megosztottátok velem a nehézségeket és örömöket egyaránt. Köszönöm kollégáimnak a Károli Pszichológiai Intézetében a megértést és támogatást, főleg Bencének, aki kiváló „research buddy” volt, remélem viszonozhatom.

Hálás vagyok a családomnak, barátaimnak azért, hogy érdeklődésükkel, türelmükkel, szeretetükkel támogattak ez alatt az öt év alatt. Köszönöm gyerekeimnek, hogy továbblendítettek az utolsó év megpróbáltatásain. Köszönöm férjemnek, Szabolcsnak, hogy szakmai értelemben is társam volt.

Köszönöm édesanyámnak, hogy végig kitartott. Sajnálom, hogy kevesebb időt tölthettem vele, mint szerettem volna. A disszertációt neki ajánlom.

1. Bevezető

Az 1980-as években, főként az angolszász országokban egy új közigazgatási irány, a közmenedzsment modell (*New Public Management*) jelent meg, mely alapvetően az üzleti világ szervezeti folyamatait alkalmazta a közszféra területein (Volacu, 2018). Miközben a döntések és a feladatok végrehajtása a központi irányítástól a korábbinál nagyobb autonómiával rendelkező szervezetekhez került, így egyre inkább decentralizálttá vált, a szolgáltatások minőségét, az üzleti gyakorlatnak megfelelően, az egységes standardok meghatározása, a teljesítmény mérése és a (költség)hatékonyság biztosította (Hood, 1991). A ráfordítások és folyamatok vizsgálata helyett, mivel azok nem minden esetben jártak együtt a nagyobb eredményességgel, a kimeneti indikátorok és az egyes reformok hatékonyságvizsgálata került előtérbe (Kertesi, 2008). A ráfordításokat és kimeneti eredményeket összevető szabályozást elszámoltathatósági rendszernek hívjuk, alapvető elemei 1) az eredmények (ma jellemzően tesztalapú) mérése-értékelése, 2) az eredmények visszajelzése, mely a fejlesztést segíti elő és 3) az értékelés eredményéhez kapcsolt ösztönzés, mely a lehető legnagyobb mértékben nyilvános és átlátható (Horn, 2010).

A Gazdasági Együttműködési és Fejlesztési Szervezet (Organisation for Economic Co-operation and Development, továbbiakban OECD) 2013-ban kiadott összegzése (OECD, 2013b) a *New Public Management* megjelenése mellett más külső körülményeket is említ, melyek az értékelés előretöréséhez vezettek. Ilyen körülmény az oktatás, mint érték a globalizált világban, az oktatás expanziója, ami szükségessé teszi a szélesebb körű értékelést, vagy a felgyorsult technológiai fejlődés, mely lehetővé teszi az eredmények gyorsabb visszacsatolását, így a fejlesztési folyamatokba történő mielőbbi beavatkozást (Szemerszki, 2014).

A megfelelő tudományos eljárásokkal bizonyított tényeken, adatokon alapuló szakpolitikai döntéshozás és az elszámoltathatóság elvének elsősorban az angolszász országokban történő elterjedése a bizonyítékokon alapuló kutatási eredmények térnyeréséhez vezetett (Commission of the European Communities, 2010). Ennek a trendnek az oktatás területén két meghatározó mérföldköve volt, mindkettő az Amerikai Egyesült Államok oktatáspolitikai döntéshozásához köthető, de a hatásaik globális szinten is jelentkeztek. Az egyik az *amerikai tudománytámogatási törvény* (Education Sciences Reform Act of 2002, 2002), mely az oktatással kapcsolatos kutatások és fejlesztések elsődleges eszközéül a kvantitatív szemléletű, azon belül is a randomizált

kísérleti elrendezésű vizsgálatokat jelölte meg. A tudománytámogatási törvénnyel kapcsolatban kritikaként fogalmazódott meg (Berliner, 2002), hogy a tudományos kutatásokat a támogatás eszközével aránytalanul a kvantitatív, gyakorlatorientált, teljesítményközpontú kutatások irányába tolja el, míg az elméleti jellegű alapkutatások hátrányba kerülnek, ami végeredményben az alkalmazott kutatások minőségi romlásához vezethet. További kritika, hogy a humán tudományok vizsgálódásainak tárgyai túlságosan bonyolultak a törvényben preferált empirikus vizsgálati módszerek lehetőségeihez képest, így ezek a kutatások a kívánt eredmény elérését (a döntéshozatal vagy az értékelés támogatását) nem tudják segíteni.

A másik jelentős állomás az amerikai *'No Child Left Behind'* törvény (NCLB Act of 2001, 2002), mely kötelezővé teszi a tanulói teljesítmények mérését és állami standardok kialakítását a lemaradó tanulók sikeres felzárkóztatásának érdekében. Eredeti célja, hogy szövetségi szinten szabályozza és írja elő a minimális oktatási színvonal elérését, különös tekintettel a veszélyeztetett tanulói csoportokra (etnikai vagy nyelvi szempontból hátrányos helyzetű, sajátos nevelési igényű tanulók) (Tomasz, 2011), ugyanakkor mind az elérendő célokat, mind a mérés rendszerét maguk az egyes szövetségi államok határozták meg. A törvény rendszeres és széles körű tanulóiteljesítmény-mérési követelményt ír elő, kijelölve azokat az iskolákat, melyeknek fejlesztésre, támogatásra van szükségük. A törvénnyel kapcsolatosan több kritika is megfogalmazódott (Dennis, 2017), ezek egyike, hogy a fókuszba került szegényebb térségekben működő iskolák nem kaptak megfelelő támogatást; a tesztek használata beszűkítette a tanórai tevékenységeket és távolította a tanárok nézeteit a tanulási tevékenység reflektív vizsgálatától; a teljesítmény küszöbértéke körüli tanulók támogatására fókuszálta a pedagógiai munkát. A NCLB törvényt 2015-ben felváltotta az *Every Student Succeeds* törvény (ESSA, 2015), mely megtartotta a tanulói teljesítmények mérésének szükségességét, de támogatást is előír a lemaradó iskoláknak (mind a tanárok, mind az eszközök tekintetében), több teret enged az államoknak egyéb értékelő eszközök használatában, és szélesíti a fejlődés és tanulás monitorozására szolgáló mérőeszközök palettáját (Dennis, 2017).

A közszféra eredményességével és elszámoltathatóságával kapcsolatos igény növekedése (Halász, 2013) következtében értékelések (*evaluation*) kapcsolódhatnak mind a döntések megalapozásához, mind a köztes állapotok ellenőrzéséhez, mind a megvalósult programok hatásának vizsgálatához. Az oktatáskutatás egy egész ágazata épült a mérés-értékelés feladatra, megrendelői és felhasználói lehetnek helyi szervezetek,

nemzetállamok vagy nagy nemzetközi szervezetek (Európai Unió, OECD) oktatáskutatással foglalkozó szegmensei is (Halász, 2013). Különösen érdekes az Európai Unió példája: mivel nemzetállamok szintjén közvetlen irányítási hatáskörrel nem rendelkezik az oktatás területén (ez nemzeti hatáskörben van), ezért a tagországok értékelésével, jó példák és gyakorlatok gyűjtésével, egységes indikátorokkal igyekszik elősegíteni a tagországok egymáshoz közelítését (Commission of the European Communities, 2010). Hasonlóan értelmezhető az OECD Programme for International Student Assessment (továbbiakban PISA) mérésének szerepe, mely a mérési területek kiválasztásával (Biesta, 2011) vagy a mérés módjának megváltoztatásával (Komatsu & Rappleye, 2017) befolyásolhatja a nemzeti oktatáspolitikai célokat. Biesta (2009) az NCLB kritikáihoz hasonló aggályokat a nemzetközi nagymintás tanulóiteljesítmény-mérésekkel kapcsolatban is felvet, miszerint a tanuló egyetlen aspektusának – bizonyos mérési területeken elért teljesítménynek – túlértékelése eltorzítja a tanítás komplex céljainak egyensúlyát.

A nemzetközi (tanulói teljesítményt vizsgáló) mérések közül Magyarországon az International Association for the Evaluation of Educational Achievement (továbbiakban IEA) Trends in International Mathematics and Science Study (továbbiakban TIMSS) (Mullis & Martin, 2017) és a Progress in International Reading Literacy Study (továbbiakban PIRLS) (Martin et al., 2016), illetve az OECD PISA (OECD, 2019b) mérések az általános iskolában és középfokon tanulók különböző korosztályait mérik a szövegértés, matematika, természettudományok és alkalmanként egyéb területeken.

Magyarországon 1986 és 1999 között a MONITOR mérés képviselte az átfogó teljesítménymérést a köznevelés területén (D. Molnár et al., 2012). Ennek alapjain, a nemzetközi mérések, elsősorban a PISA mintájára és annak metodikáját követve hozták létre a 2000-es évek elején az Országos kompetenciamérést (továbbiakban OKM). Ennek célja egyrészt, hogy meghonosítsa Magyarországon a nemzetközi mérési kultúrát a köznevelésben, másrészt, hogy visszajelző rendszer legyen a köznevelési intézmények és az oktatásirányítás felé. Az OKM alapkoncepciója szerint a mérési egység ideálisan az iskola, mivel itt tud leggyorsabban hasznosulni a visszajelzés, illetve a tanulói szintű mérési hibák is ezen a szinten egyenlítődnek ki. A korszerű nevelésszociológiai álláspont szerint is elmondható, hogy a tanulói eredményesség és a pedagógus eredményessége a szervezet eredményességéből fakad, így az oktatásfejlesztés legfontosabb fókusza az iskolai szervezet (Nahalka & Sipos, 2016).

A tanulók teljesítményét mérő hazai (OKM) és nemzetközi (PIRLS, TIMSS, PISA) tesztek mindegyike eredetileg papír-ceruza alapú teszt. A három nemzetközi mérés közül kettő (PIRLS és TIMSS) ma oly módon digitális, hogy számítógépen történik a kitöltés, de továbbra is előre rögzített tesztfüzetek feladatait oldják meg a diákok. Ezek a mérések a feladatok írása és a mérés szervezése tekintetében használják ki a technológiában rejlő lehetőségeket, ebben az értelemben a számítógépes tesztelés első szintjén (Csapó et al., 2008) állnak. A tesztszerkesztés esetében az adaptív tesztelés egy továbblépési lehetőség (Magyar, 2012). Az adaptív mérés olyan eljárás, mely során a teszt felvétele közben a következő kérdés (számítógépes adaptív tesztelés, *computerized adaptive testing*, továbbiakban CAT) vagy kérdéscsoport (többszakaszos adaptív tesztelés, *multistage adaptive testing*, továbbiakban MST) kiválasztása a korábbi válaszok függvényében történik (Mead, 2006; Weiss & Kingsbury, 1984). A PISA mérés 2018-ban többszakaszos adaptív tesztet alkalmazott a fő mérési területen (szövegértés) (ld. 3.2 fejezet). A hazai kompetenciamérés 2021-ben még papír-ceruza alapon történt, a 2022. és 2023. évi OKM során már számítógépen dolgoztak a tanulók, de ezekben a mérési körökben a PIRLS és TIMSS mérésekhez hasonlóan még digitalizált feladatlap-változatokat töltöttek ki.

Az adaptív mérések módszere (ld. 2.3 fejezet) már több évtizede rendelkezésre áll (Weiss & Kingsbury, 1984; Weiss, 2011; Chang, 2015). Mivel hagyományos, tehát számítógép alkalmazását nélkülöző tesztfelvétel esetén minden tesztalany mellé felmérésvezetőre van szükség, ezért költségessége miatt e módszer nem terjedt el. Az informatikai fejlődés következtében lehetőség nyílt a mérések automatizálására, egyéni értékelésre és gyorsabb visszajelzésre (Szemerszki, 2014), ezért egyre több esetben valósulnak meg adaptív mérések, lásd például a pszichológia (Gonthier et al., 2018), az orvoslás (Petersen et al., 2013) vagy az oktatás (Nogami & Hayashi, 2010; Wang et al., 2019) területein. Magyarországon a Szegedi Tudományegyetemen folynak adaptív teszteléssel, elsősorban többszakaszos adaptív tesztekkel kapcsolatos kutatások (Csapó et al., 2008; Magyar, 2015; Molnár, 2010) (ld. 3.4.2 fejezet).

Az adaptív mérések, sőt maguk a számítógépes mérések fejlesztése papír-ceruza tesztek alapján a financiális és szervezési kérdéseken túl módszertani problémákat is felvet. A mérési módszertan megváltoztatása esetén mindig kérdéses mind a validitás, mind a reliabilitás (Nagybányai-Nagy, 2006b, 2006a). Validitáson azt értjük, hogy a mérés által mért jelenség (konstruktum) továbbra is megfelel annak a jelenségnek, amelyet a papír-ceruza teszt mért. Reliabilitáson pedig egyrészt azt értjük, hogy a teszten

a vizsgálati alany a különböző tesztfelvételein hasonló eredményeket ért el, másrészt a papír-ceruza és a számítógépes vagy (számítógépes) adaptív teszt eredménye nem tér el jelentősen (ld. 2.5 fejezet).

Első fejlesztési lépésnek tekinthető a papír-ceruza teszt számítógépes mérőeszközre történő cseréje. Az adaptív tesztelésnek előfeltétele a számítógépes adatfelvétel, azonban adaptív módszertan alkalmazása nélkül is elképzelhető a tesztmédium cseréje, erre több nemzetközi példa is van a tanulóiteljesítmény-mérések körében (PISA 2015, TIMSS 2019, PIRLS 2022) (ld. 3.1 fejezet). A fejlesztésnek ezen a pontján a validitási kérdés, hogy az új médiumon mért konstruktum továbbra is megfelel-e a papír-ceruza teszttel mért jelenségnek. Ez leginkább az IKT eszközök használatával kapcsolatos készségek, mint további mérési dimenzió problémáját érinti. A reliabilitás problematikája pedig az, hogy a két módon mért pontszám megegyezik-e.

Második fejlesztési lépés – feltéve, hogy az első lépés problémái megfelelően kontrolláltak – a számítógépes mérés során a lineáris teszt cseréje az adaptív tesztre. Ebben az esetben a validitási kérdésnek több része is van. Egyrészt az eltérő tesztutak miatt akár minden tesztalany más-más tesztet tölthet ki, tehát kérdéses lehet, hogy mindenki esetében ugyanazt a jelenséget mérjük-e. Ezt a problémát orvosolja, hogy a feladatbank tételei ugyanannak a jelenségnek a mérésére készülnek, illetve több alterület kombinációja esetén az adaptív tesztek is képesek a tartalmi területek arányának kontrolljára. Szintén probléma lehet, hogy számítógépes adaptív teszt esetében kizárólag automatikus kiértékelésű itemek szerepelhetnek, az aktuális képességbecsléshez, mely elengedhetetlen része a folyamatnak, csak ezeket lehet felhasználni. Lehetséges megoldás, hogy nyílt végű itemeket (ahol az önállóan konstruált válaszokat képzett kódolók értékelik) szintén tartalmaz a teszt, de ezek a tesztutak kialakításában nem vehetnek részt, a becslést csak a mérés után módosíthatják (OECD, 2019d). A nyílt végű itemek általában részei a tanulóiteljesítmény-méréseknek, a tartalmi keretek szerint bizonyos gondolkodási műveletek leginkább ilyen formában mérhetők (Mullis et al., 2021; Mullis & Martin, 2019; OECD, 2019a). Elhagyásuk szintén a mért konstruktumok különbözőségének vizsgálatát teszi szükségessé.

Az Országos kompetenciamérés hazai viszonylatban széles körben használt mérőeszköz a köznevelléssel kapcsolatos kutatások területén. A PISA mérés nyomán (OECD, 2019d; Yamamoto, Shin et al., 2018) az OKM esetében is lehetséges fejlesztési irány, hogy adaptívan valósuljon meg. Egy ilyen fejlesztés számos előnnyel járhat. A számítógépes mérés esetén lehetséges az eredmény azonnali, de legalább gyorsabb

visszajelzése a tanulók felé, a mérés teljes feldolgozásának ideje lerövidül, a költségek mérséklődhetnek (Balázsi et al., 2021; Szemerszki, 2014). Az adaptív mérés esetén kevesebb a tanulók által megoldandó itemek száma, ami rövidebb tesztidőt és így kisebb megterhelést jelent (Frey & Seitz, 2009), valamint legalább ugyanolyan pontos képességbecslést biztosít (Csapó et al., 2008), ami leginkább a képességskála szélén elhelyezkedő tanulók esetében hozhat a nem adaptív módszerrel végrehajtott méréshez képest jobb eredményeket.

Disszertációm fókuszában annak vizsgálata áll, hogy az OKM esetében milyen feltételei lennének a szakmai és mérésmódszertani szempontból sikeres papír-ceruza teszt, számítógépes mérés, adaptív számítógépes adatfelvétel közötti átmenetnek. A mérés módszertanának ilyen szintű megváltoztatása – papír-ceruza formáról adaptív számítógépes környezetre történő átültetése – körültekintő előkészítést igényel és számos következménnyel jár. Egy OKM-hez hasonló trendvizsgálat alapja, hogy az új, adaptív módszer bevezetése előtti és utáni eredmények egymással összehasonlíthatók legyenek. Az ehhez szükséges előzetes kutatások, vizsgálatok elvégzésére nemzetközi mérések esetében vannak példák (ld. 6.1 fejezet). Magyarországon ilyen előzetes vizsgálat publikációja az OKM esetében eddig nem történt, így a disszertációban az adaptív mérésre vonatkozó kutatás (ld. 6.4 fejezet) úttörő ebben a tekintetben. Fontos cél továbbá az adaptív mérési technológia további megismertetése a hazai szakmai körökben, mely célkitűzés nem mellesleg összhangban van az OKM egyik eredeti céljával, a nemzetközi mérési kultúra megismertetésével (Balázsi et al., 2005; Csíkos & Vidákovich, 2012).

A hazai szakirodalom kevéssé foglalkozik a tanulói mérések digitalizálásával és az adaptív mérésekkel – eltekintve a Szegedi Tudományegyetemen folyó munkától. Magyar vizsgálata (2015) az eDia mérés-értékelési rendszerben megvalósuló mérésre irányul, és a többszakaszos adaptív tesztelést (MST) vizsgálja valódi adatfelvétel alapján. Ezzel szemben kutatásomban a számítógépes adaptív tesztelésre (computerised adaptive testing, CAT) alkalmaztam szimulációs módszert, ami újdonság a korábbi hazai vizsgálatokhoz képest. A szimulációs vizsgálat előfeltevések vagy a korábbi mérés(ek)ből származó ismeretek alapján, valódi vagy a valószínűségi modellnek megfelelő imitált tesztkitöltést használ fel, és egy lehetséges adaptív teszt alapján kiszámítja a becsült tesztpontszámot (Sari, 2020). Ezt követően az előre meghatározott képességfejlettség és a becsült képességpont alapján következtethetünk a leendő mérés tulajdonságaira. Esetemben a szimuláció alapját az OKM papír-ceruza méréseiből származó itemek jellemzői és tanulói adatok alkotják. A kutatás teljes folyamata

mintaként szolgálhat más papír-ceruza mérések számítógépes méréssé tétele és főképpen számítógépes mérések adaptív mérésre történő átültetésének előkészítéséhez, akár a tartalmi validitás ellenőrzését, akár az adaptív elemek kiválasztását illetően.

2. Méréselméleti háttér áttekintése

A mérés azzal foglalkozik, hogy a tudományok (akár természettudomány, akár humán tudományok) területén a jelenségekhez bizonyos szabályok szerint matematikai objektumokat (pl. számokat) rendeljen (Stevens, 1946), és az adott tudomány a számok egymáshoz való kapcsolata alapján feltárja a mért jelenségek tulajdonságait (Nahalka, 2018). A méréselméletek ezzel szemben nem magával a méréssel, hanem a lehetséges mérések és hozzárendelési szabályok tulajdonságaival foglalkoznak. Míg a klasszikus méréselmélet a mennyiségek összeadódását vagy kiegyenlítődését tekintette alapvető tulajdonságnak (von Helmholtz, 1977), addig a skálák elmélete a mennyiségek közötti relációkból indul ki (Stevens, 1946). A két elméletnek a reprezentációs méréselmélet egyfajta szintézise, mely egyetlen struktúrába tömöríti adott halmaz elemeit, a köztük értelmezett relációkat és az elemekre és relációkra vonatkozó axiómákat (Nahalka, 2018).

Tesztnek nevezzük azt a mérőeszközt, amely egyértelműen kiértékelhető feladatokból (item) áll, azaz bármely válaszról eldönthető, hogy helyes vagy helytelen (dichotóm item) (Kontra, 2011), vagy a maximális pontszámból hány pontot ér (több pontos item). A klasszikus tesztelméletben a teszt eredménye az összpontszám vagy annak valamilyen transzformációja (pl. százalékos megoldottság), míg a modern tesztelméletben ez a válaszmintázat alapján valószínűségi modellekkel (ld. 2.2 fejezet) számított képességpont. Mivel a teszt, és általában a mérőeszköz célja, hogy a jelenségre és a vizsgálati személyre vonatkozó megállapításokat tegyünk, ezért szükséges a mérőeszköz alkalmasságának vizsgálata.

A tesztek esetében többféle tulajdonság vizsgálata szokásos. Ilyen tulajdonság az *érvényesség vagy validitás*, amely azt jelenti, hogy valóban azt a jelenséget mérjük, amelyre a megállapítást tenni szeretnénk (Nagybányai-Nagy, 2006b). A *megbízhatóság vagy reliabilitás* ezzel szemben azt vizsgálja, hogy amit mérünk, bármi legyen is az, pontosan mérjük-e (Nagybányai-Nagy, 2006a). Ezen a ponton megkülönböztetünk ismétléses reliabilitást, vagyis különböző időpontban felvett tesztek, illetve belső konzisztenciát, vagyis egyazon teszt tetszőleges részeinek, akár itemeinek egymással való kapcsolatát. Világos tehát, hogy nincs érvényesség megbízhatóság nélkül. A reliabilitás ugyanakkor felfogható úgy is, hogy egy item-együttes mennyire sikeresen tárja fel az egyéni különbségeket (Cronbach, 1951 idézi Nagybányai-Nagy, 2006a), ebben az értelemben tehát az itemek mellett a mért mintát is jellemzi.

A következő fejezetekben röviden bemutatok néhány lehetséges reliabilitási mutatót, áttekintem a jelenlegi nagymintás tanulóiteljesítmény-mérések által alkalmazott modern tesztelmélet jellemzőit, a modern tesztelméletre és a számítógépes felmérésvezetésre épülő adaptív tesztelési eljárásokat, végül a papír-ceruza és a számítógépes mérés közötti különbséget, a médiahatást.

2.1. Tesztek megbízhatósága[1]

Egy teszt eredménye a klasszikus tesztelméletben a tesztalany képességfejlettsége és a mérőeszközből származó véletlen hiba összege. A modern tesztelméletben ehhez képest a képességfejlettség valamilyen véletlen függvény szerepel, de a véletlen hiba itt is megjelenik. A reliabilitás mutatók jellemzően a teszteredmények közötti kapcsolatot, együttjárást (korrelációt) mérik (Nagybányai-Nagy, 2006a), így a magasabb érték azt jelzi, hogy a véletlen hiba mértéke elmarad a képességfejlettségből származó résztől. Ugyanakkor az alacsony érték alacsony konzisztenciából vagy a minta homogén voltából is következhet.

A *belső konzisztencia*, azaz az itemek együttjárásának ellenőrzésére az adott teszt feldarabolása és a tesztrészek összehasonlítása adott lehetőséget (felezéses módszerek, pontbiszeriális korreláció), azonban a felosztás mikéntjétől független mutató először Kuder és Richardson eredeti mutatója (KR-20) (Kuder & Richardson, 1937) volt (ld. még 6.3 fejezet). A formula dichotóm itemek esetében működik, az itemek megoldottsága alapján, lényegében egy átlagos kitöltő hibamutatójára alapoz, ezzel a szélsőséges kitöltők hibáját alulbecsli (Wright & Stone, 1999). A formulát Cronbach (Cronbach, 1947, 1951) többpontos itemek esetére is általánosította, ez a ma is ismert Cronbach-alfa mutató, amely manapság a lineáris tesztek esetében az egyik legfontosabb, elsők között kiszámolt jellemző. A mutató az itemek közötti kovarianciákat veti össze a teljes teszt eredményének varianciájával.

A Cronbach-alfa mutatónak több korrigált változata (Cho, 2016) és alternatívája van. Ilyen alternatíva többek között a McDonald féle omega (McDonald, 1999), mely a megerősítő faktoranalízis faktor-töltéseinek vizsgálatára alapul. A két mutató összehasonlítása jellemzően az omega használatát javasolja (pl. Hayes & Coutts, 2020;

[1] A fejezet részben a Magyar Pszichológiai Szemlében megjelent cikk (T. Kárász et al., 2022) alapján készült.

Malkewitz et al., 2023; Peters, 2014), azonban a szükséges eljárás a legtöbb statisztikai programnak nem része, ezért alkalmazása nehézkes lehet.

A Cronbach-alfa mutatónak több kritikája ismeretes, azonban ezek egy része inkább a helytelen használatból következik. A mutató használatának előfeltétele, hogy a mutatók egyetlen jelenséget mérjenek, azonban visszafelé az állítás nem igaz, a mutató magas értéke nem igazolja az itemek egy faktorba tartozását (pl. Schmitt, 1996). Hasonló alkalmazási hiba, hogy a Cronbach-alfa magas értékét elvárva a 0,7 (esetleg 0,8) küszöbérték elérését tartják kívánatosnak (Peters, 2014), holott a mutató értéke mesterségesen növelhető a teszt hosszának növelésével, hasonló itemek alkalmazásával vagy finomabban mérő itemtípus (dichotóm helyett pl. Likert-skálák) használatával (Nagybányai-Nagy, 2006a). A küszöbértékkel kapcsolatos további probléma, hogy érzékeny magának a jelenségnek a jellegére, így pl. intelligenciateszteknél a 0,8–0,9-es érték könnyen elérhető, míg a változékonyabb jelenséget mérő attitűdskálák esetében 0,5 körüli értékek elfogadhatók (Horváth, 1997). Éppen ezért javasolt lehet a kívánatos Cronbach-alfa értéket a jelenség stabilitása, az itemek száma és finomsága, valamint az itemek között elvárt együttjárás alapján meghatározni (T. Kárász et al., 2022).

A modern tesztelmélet, azon belül a Rasch-modell (ld. 2.2 fejezet) esetén a személyszeparációs mutató (Wright & Masters, 1982) is alkalmazható (ld. még 6.3 fejezet). Maga a Rasch-modell a képességpont becsléséhez szolgáltat egy mérési hibát (standard error, SE), így a személyszeparációs mutató a belső konzisztencia másik aspektusára, a tesztalanyok megkülönböztető képességére koncentrál. A három mutatót (KR-20, Cronbach-alfa, személyszeparációs mutató) összevetve Anselmi és munkatársai (2019) szimulációs vizsgálatukban úgy találták, hogy szimmetrikus teljesítmény-eloszlás esetén a mutatók hasonló eredményre vezetnek, ferde eloszlások esetében a személyszeparációs mutató bizonyul a legkonzervatívabbnak, azaz a legkevésbé becsüli felül a teszt reliabilitását.

Adaptív tesztek esetében a Cronbach-alfától eltérő megbízhatósági mutatókra van szükség, mert a teszt nem azonos minden kitöltő esetében. Mivel mindenki a számára leginkább megfelelő itemet kapja az adott, korábbi válaszainak megfelelően, ezért egyedi tesztutak keletkeznek, sőt minden kitöltőnek egyedi megbízhatóság számítható. Elképzelhető, hogy bizonyos itemek egyetlen tesztváltozatban sem szerepelnek együtt, tehát a közöttük értelmezett kovariancia sem értelmezhető. Ennek vizsgálatához olyan mutatóra van szükség, mely erre a sajátosságra is fel van készítve, ugyanakkor kiállja a matematikai általánosítás próbáját. Ilyen mutató lehet a hagyományosan használt

Cronbach-alfa egy igen speciális esete, a KR-20, mely mindössze a dichotóm itemek megoldottságát használja.

2.2. A modern tesztelméletről (IRT)

Mivel az OKM és a módszertani példaképnek tekintett PISA felmérés a képességpontokat és az itemek[2] jellemzőit a modern tesztelmélet (Item Response Theory, továbbiakban IRT) alapján számítja, valamint az adaptív tesztek mechanizmusa szintén valamilyen IRT modellre épül, ezért szükséges ezen háttér rövid bemutatása.

Legyen adott egy minta, melynek tagjai egy feladatbank itemeit oldják meg. Az itemek összessége jellemezze a vizsgált jelenséget, melyet nevezzünk képességnek. Ekkor a helyes feladatmegoldás valószínűsége a vizsgált személy képességfejlettségének objektív mértékétől, valamint az adott feladat objektív nehézségétől függ (Rasch, 1960, idézi Nahalka, 2018). Az itemek sikeres megoldásának valószínűségét elsősorban az item nehézségével jellemezzük. A populáció tagjainak sikeres feladatmegoldási valószínűségét pedig a képességfejlettséggel jellemezzük.

A pszichometriában a képességfejlettség becslésére egypontos itemek esetén több IRT modell (egyparaméteres Rasch-modell, két-, illetve háromparaméteres modellek) is ismert és használatos (Auxné Bánfi et al., 2014; Lannert, 2015). A j item megoldási valószínűsége a θ képességfejlettség esetén az egyparaméteres (1), kétparaméteres (2) és háromparaméteres (3) modell egyenlete alapján számítható, ahol az itemparamétereket a és a_j (meredekség), b_j (nehézség) és c_j (tippelési paraméter) jelölik (DuToit, 2003).

(1) $$P_{1j}(\theta) = \frac{1}{1+e^{-a\left(\theta - b_j\right)}}$$

(2) $$P_{2j}(\theta) = \frac{1}{1+e^{-a_j\left(\theta - b_j\right)}}$$

(3) $$P_{3j}(\theta) = c_j + (1 - c_j)\left(\frac{1}{1+e^{-a_j\left(\theta - b_j\right)}}\right)$$

Adott képlettel egy θ képességfejlettségű személy és a j item paraméterei alapján egyértelműen meghatározható a sikeres feladatmegoldás valószínűsége. Ismert paraméterekkel rendelkező itemekből álló teszt esetén a képességfejlettség becslése a

[2] Itemnek nevezem a feladat azon egységét, mely a pontozáskor egy egységet képez. Ebben az értelemben egy *a* és *b* részből álló feladat két itemet tartalmaz, ugyanakkor négy igaz/hamis kérdést tartalmazó feladat egy item, négy részitemmel.

képességskála azon értéke, amelyre az itemek paramétereivel számított megoldási valószínűségekkel az itemeken elért válaszmintázat a legvalószínűbb.

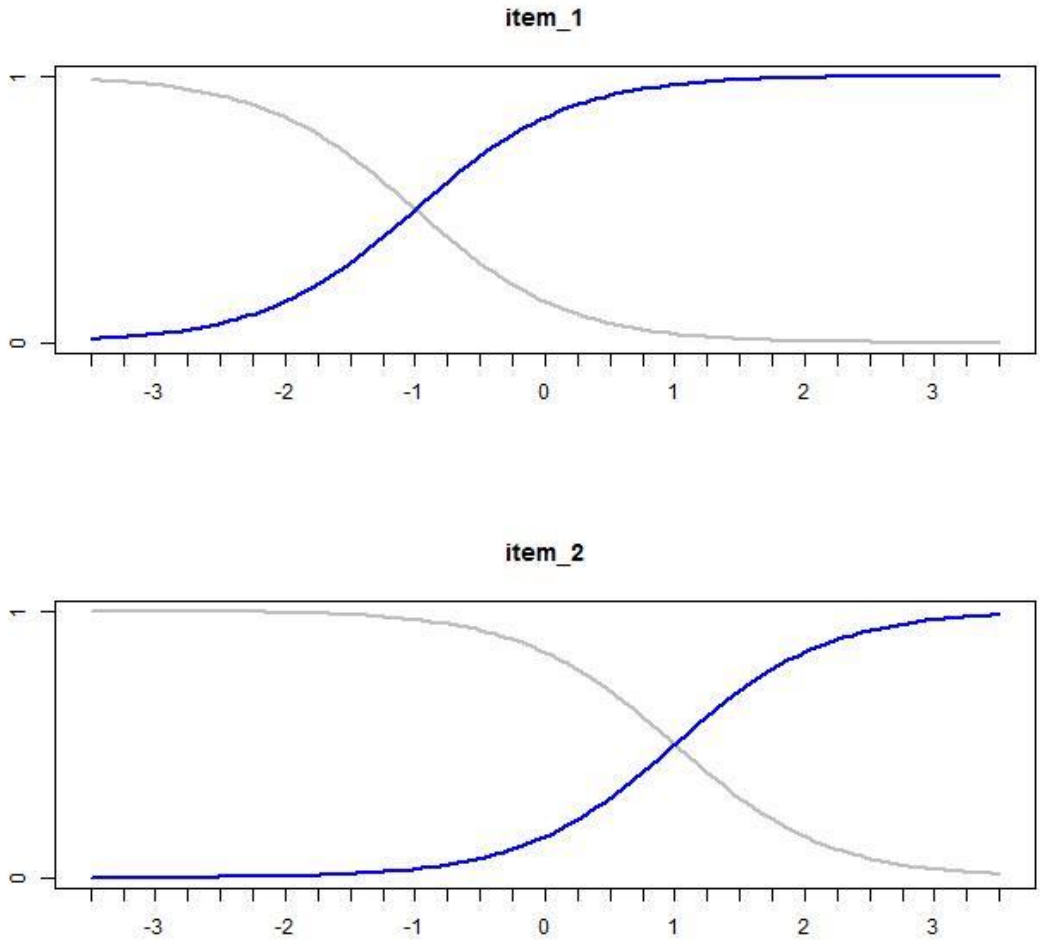
A képességfejlettség és az item nehézségének közös skálája azt jelenti, hogy értelmezhető az a kijelentés, hogy egy válaszadó képességpontja megegyezik egy item nehézségével, vagy kisebb/nagyobb annál. A megoldási valószínűség további tulajdonsága, hogy tesztfüggetlen, azaz ugyanazon képességfejlettséget mérő hasonló (pl. a képességskálához azonos paraméterekkel rögzített közös itemeket tartalmazó) tesztek adott kitöltőnek hasonló képességpontot számítanak, és mintafüggetlen, azaz ugyanazt a tesztet hasonló (pl. a képességskálához azonos képességfejlettséggel rögzített közös tesztkitöltőket tartalmazó) részpopulációkkal kitöltetve az itemeknek hasonló nehézséget számítunk. Az IRT modellek egy része tipikusan olyan itemekre megfelelők, amelyek eredménye 0 vagy 1 pont lehet, de léteznek többpontos itemekre alkalmazható modellek is. Az itemek jellemzőivel kapcsolatos fontosabb fogalmak az alábbiak.

Karakterisztikus görbe: az a képességskálán értelmezett függvény, mely adott item paraméterei (nehézsége, meredeksége, tippelési paramétere) alapján meghatározza, hogy adott képességfejlettség mellett mekkora jó megoldás valószínűsége (Molnár, 2006). A j. item karakterisztikus görbéje a $\theta \rightarrow P_j(\theta)$ hozzárendelés ábrázolása, vagyis a karakterisztikus görbe egyes pontjai a vízszintes tengelyen a képességskála egy képességpontját (θ), a függőleges tengelyen a jó válasz valószínűségét jelölik ki ezen képességpont esetén $(P_j(\theta))$.

Az item nehézsége és a válaszadó képességpontja közötti kapcsolat: az item nehézsége a képességskála azon értéke, amelyben a karakterisztikus görbe meredeksége maximális. Ez éppen a görbe inflexiós pontja. A fentiekből következik, hogy a két fogalom egyszerre, egymás által is értelmezhető. Egy adott item nehézsége a képességskála azon pontja, amely képességfejlettséggel rendelkező válaszadók éppen 50% eséllyel válaszolják meg jól a kérdést (feltéve, hogy nincs lehetőség tippelésre). (Többpontos item esetében a képességskála azon pontja, ahol a 0 és a maximális pontszámot érő válasz esélye éppen ugyanannyi.) Adott válaszadó képességfejlettsége megfelel azon itemek nehézségének, melyeket éppen 50% eséllyel old meg az adott személy. Az 1. ábra olyan itemek karakterisztikus görbéit mutatja be, melyek nehézségükben különböznek: a felső egy könnyebb itemhez, az alsó egy nehezebb itemhez tartozik.

1. ábra

Könnyebb item (felül) és nehezebb item (alul) karakterisztikus görbéje (Forrás: saját ábra)



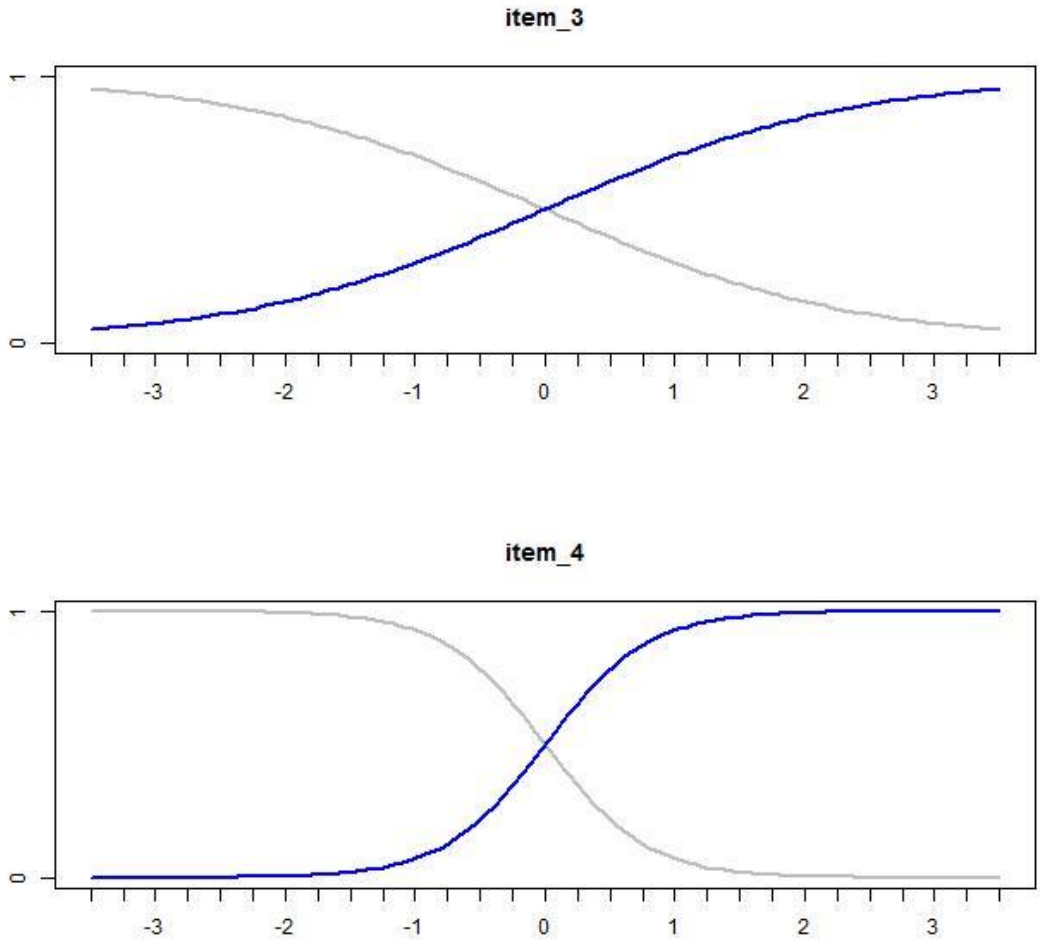
Megjegyzés. A vízszintes tengelyen a képességfejlettséget, függőleges tengelyen az item megoldásának valószínűségét ábrázoljuk, a helyes megoldás valószínűségét lila színnel, a sikertelen megoldás valószínűségét szürke színnel. A két item meredeksége megegyezik (a = 1), de a nehézsége különböző (b_1 = -1 és b_2 = 1).

Meredekség: (kétparaméteres modell) a görbe maximális meredeksége. Ezt éppen a karakterisztikus görbe inflexiós pontjában veszi fel. A nehézség és a képességfejlettség kapcsolatából (a megoldási valószínűség egyenletéből) következik, hogy a képességfejlettség növekedésével a jó megoldás valószínűsége is nő. A képességfejlettség oldaláról nézve a meredekség azt mutatja meg, hogy a megoldás valószínűsége milyen gyorsan nő vagy csökken a képességskálán a nehézségtől távolodva (Auxné Bánfi et al., 2014, p.33). Az itemek szemszögéből azt jelenti, hogy a nehézség mekkora képességpont-környezetében mér az item, avagy milyen a diszkrimináló tulajdonsága (a válaszadók között) (Molnár, 2006). Utóbbi két tulajdonság ellentétes irányú a következő értelemben: egy olyan item, amely esetében a nehézségtől távolodva gyorsan változik a megoldás valószínűsége (magas a meredekség), jó diszkrimináló képességgel rendelkezik, azaz jól szétválasztja a válaszadókat a képességfejlettségük szerint, azonban a képességskálának csak kis részén (a nehézség környékén), azaz lényegében nem különbözteti meg egymástól azokat a válaszadókat, akiknek a képességfejlettsége hasonló, de a nehézségtől távolabb esnek. Hasonlóan, egy olyan item, amely esetében a nehézségtől távolodva lassan változik a megoldás valószínűsége (alacsony a meredekség), gyenge diszkrimináló képességgel rendelkezik, azaz nem választja szét a válaszadókat a képességfejlettségük szerint, azonban a képességskálának nagyobb részén (a nehézségtől távolabb is) mér, azaz azokról a válaszadókról is információt ad, akik a nehézségtől távolabbi képességfejlettséggel rendelkeznek. A 2. ábra felül egy kisebb, alul egy nagyobb meredekségű item karakterisztikus görbéjét mutatja be.

Tippelési (guessing) paraméter: (háromparaméteres modell) Tipikusan feleletválasztós itemek esetében igaz, hogy a megoldás valószínűsége a képességskála legalsó szintjén sem 0, mivel tippeléssel a válaszadó ráhibázhat a jó megoldásra. Négy nagyon hasonló válaszlehetőség esetén ez a valószínűség 25%, de nyilvánvalóan rossz válaszlehetőségek esetén magasabb is lehet. (Ha a teszt kitöltői nagy arányban ki tudják zárni az egyik válaszlehetőséget, akkor kevesebb egyformán valószínű lehetőség közül tudnak tippelni.) A tippelési paraméter a karakterisztikus görbe korrekciója ezzel a valószínűségi tényezővel (*Országos kompetenciamérés—Technikai leírás,* 2010, 33.o), hogy az item modellje jobban illeszkedjen a képességskála egyes szakaszain tapasztalt megoldási valószínűséghez. Természetesen nem kötelező nullától különböző tippelési paraméterrel kiegészíteni a feleletválasztós itemek modelljét, ha az nem javítja az illeszkedést. A 3. ábra felül egy tippelési paraméter nélküli, alul egy tippelési paraméterrel kiegészített item karakterisztikus görbéit mutatja be.

2. ábra

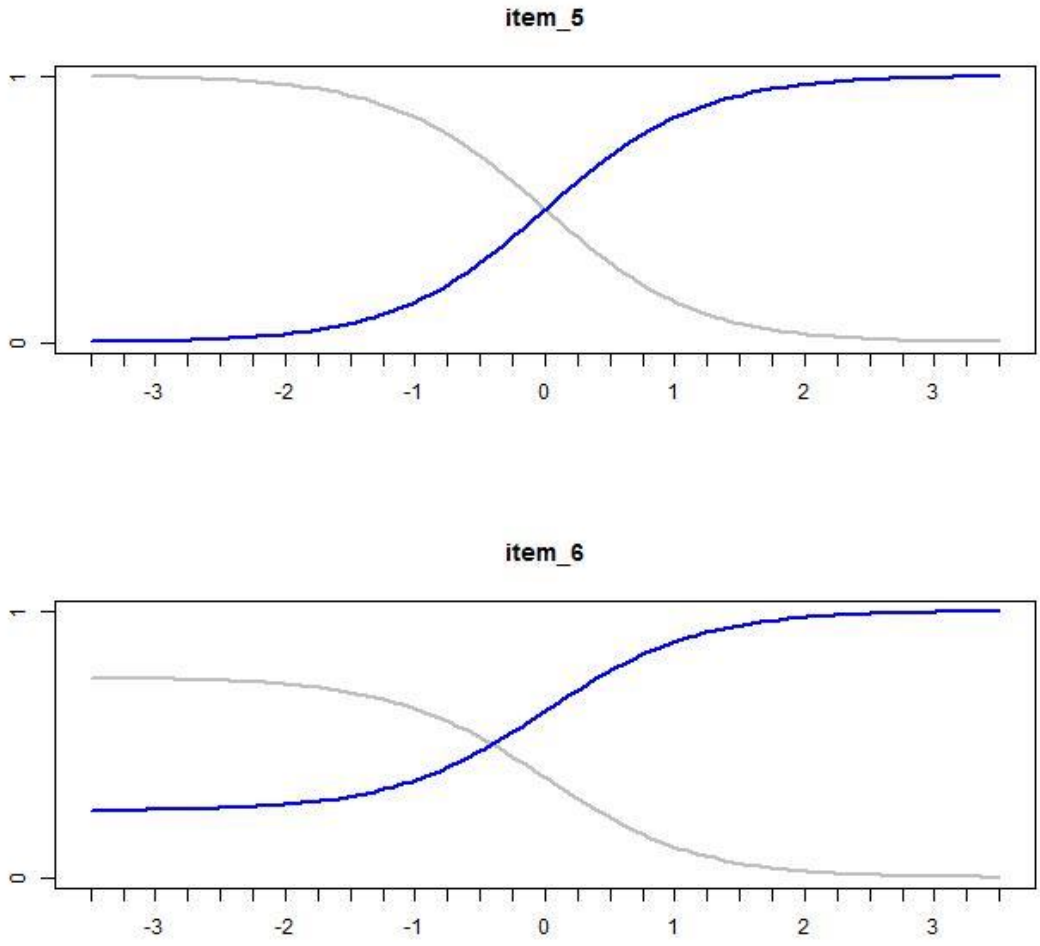
Kevésbé diszkrimináló item (item_3) és jobban diszkrimináló item (item_4) karakterisztikus görbéje (Forrás: saját ábra)



Megjegyzés. A vízszintes tengelyen a képességfejlettséget, függőleges tengelyen az item megoldásának valószínűségét ábrázoljuk, a helyes megoldás valószínűségét lila színnel, a sikertelen megoldás valószínűségét szürke színnel. A két item nehézsége megegyezik (b_3=0 és b_4=0), ami nem követelmény, hanem a szemléletesebb ábrázolást segíti, a meredeksége azonban különbözik (a_3=0,5 és a_4=1,5).

3. ábra

Tippelési paramétert nem igénylő item (item_5) és tippelési paramétert igénylő item (item_6) karakterisztikus görbéje (Forrás: saját ábra)



Megjegyzés. A vízszintes tengelyen a képességfejlettséget, függőleges tengelyen az item megoldásának valószínűségét ábrázoljuk, a helyes megoldás valószínűségét lila színnel, a sikertelen megoldás valószínűségét szürke színnel. A két item nehézsége ($b_5 = b_6 = 0$) és meredeksége ($a_5 = a_6 = 1$) megegyezik, ami nem követelmény, hanem a szemléletesebb ábrázolást segíti. Az alsó item modelljét 25%-os tippelési paraméter ($c_5 = 0$ és $c_6 = 0{,}25$) egészíti ki.

Molnár (2006) felvet néhány lehetséges problémát a Rasch-modellel és általában az IRT eljárásokkal kapcsolatban.

1) *Találgatás*, melyre feleletválasztós itemek esetén a háromparaméteres modell megoldást kínál. Számítógépes teszteknél megoldás lehet az olvasási és válaszadási sebesség figyelése (Chalhoub-Deville, 1999; Csányi & Molnár, 2021).
2) *Itemfüggőség*, azaz a feladatok egymásra épülnek vagy nagyon hasonlóak (erős függés), vagy a közös kontextus köti össze őket (gyenge függés). Ez általában tesztelési probléma, az itemek függetlenségét minden tesztben elvárjuk. Az OKM esetében az erős függés kizáró tényező, azonos feladat részei is egymástól függetlenül kell, hogy megoldhatók legyenek. A gyenge függés megengedett, ami a szövegértés teszt esetén érhető tetten. Számítógépes kérdések esetén, amennyiben a teszt rögzített, ez megoldható, változó teszt esetében azonban az erős függés kivédésére az itembankban (ld. 2.3.1 fejezet) hasonló itemek kupacait kell megjelölni, melyekből nem kerülhetnek egymáshoz közel vagy ugyanabba a tesztbe itemek.
3) *Eltérő itemműködés*, azaz a mintapopuláció részpopulációin az item nem ugyanolyan karakterisztikus görbét mutat, más a válaszadási valószínűség mintázata. Ilyenek lehetnek a sporttal kapcsolatos itemek a nemek között, illetve a vonatmenetrendre vonatkozó kérdések vidéki-fővárosi válaszadók körében. Ez szintén általános tesztfejlesztési probléma, és az OKM vonatkozásában vizsgálható többek között a halmozottan hátrányos helyzetű tanulók eltérő feladatmegoldása (Kispál & Gergely, 2022).
4) *Többdimenziósság* (többdimenzionalitás), azaz egy item nem csak egyféle kompetenciát mér. Bár az OKM igyekszik a feladatok kiválasztásánál ügyelni arra, hogy a matematika itemek ne kívánjanak magas szövegértési képességfejlettséget, de a szöveges megfogalmazás miatt mégis szükség van rá egy bizonyos szinten. A tesztfejlesztés szempontjából ez azt jelenti, hogy azok az itemek, amelyek megoldásában jelentősebb a szövegértés komponens, nem fognak jól illeszkedni az OKM matematika eszköztudás terület itemei közé, ezért a próbamérés után nem kerülnek be a főmérésbe (ld. 3.1 fejezet). Megoldás lehet a Multidimensional Item Response Theory (továbbiakban MIRT) (Hartig & Höhler, 2009), amelynek megfelelően kifejlesztett modellei azt teszik lehetővé, hogy az itemeket – és a válaszadók képességfejlettségét – egyszerre több

képességskálán is értelmezzük/mérjük. Egy ilyen modell lehetővé tenné, hogy az OKM itemeit egyszerre vizsgáljuk a szövegértés és a matematika eszköztudás irányából, azonban jelen munka keretein belül ezzel az iránnyal nem foglalkozom mélyebben, csak és kizárólag az egydimenziós jelenségek mérését vizsgálom.

2.3. Adaptív tesztelés

Az *adaptív tesztelés* olyan eljárás, melynek során a teszt következő egységét a válaszoló korábbi feleletei alapján választják ki (Luecht & Nungester, 1998; Weiss & Kingsbury, 1984). Az első ilyen teszt a Stanford-Binet-féle intelligencia-teszt volt (Weiss, 2011), ahol egy adott mentális kornak megfelelő feladatok képezik a teszt egységeit, a kérdezőbiztos pedig a teszt eredménye alapján vezeti a kérdezést. Az aktuális tesztrészre adott válaszokat képzett kérdezőbiztosok pontozzák vagy kiértékelik, és ennek megfelelően választják a teszt következő, az eredménynek megfelelő egységet. A számítógépek fejlődésével és az IRT modellek széleskörű használatával az adaptív tesztelés egyébként költséges eljárása is elterjedhetett, mivel az eredmények értékelése és a következő tesztrész kiválasztása automatizáltan, közvetlen emberi beavatkozás nélkül történhetett.

Az adaptív teszt szinonimaként használatos még a személyre szabott tesztelés (*tailored testing*) kifejezés is, elsősorban a témával foglalkozó régebbi publikációkban (pl. Kent & Albanese, 1987; Spineti & Hambleton, 1977; Warm, 1989) és az ausztrál National Program – Literacy and Numeracy (továbbiakban NAPLAN) mérés esetében (ACARA, 2020; Thompson, 2017). Az adaptív méréseknek két fő megvalósítási formája van: a számítógépes és a többszakaszos adaptív tesztelés, azonban az adaptív teszt, definíciójából fakadóan, mindenképpen tartalmaz legalább egy olyan pontot, ahol a korábbi válaszok kiértékelése és a becsült képességfejlettségnek megfelelő nehézségű következő egység kiválasztása történik.[3]

Mint látni fogjuk, a nemzetközi és hazai mérések a papír-ceruza tesztekről a számítógépes mérésre, majd kisebb-nagyobb mértékben az adaptív tesztelésre tértek át (ld. 3.1 és 3.2 fejezetek). Alább részletesen ismertetem az adaptív tesztelés két megvalósítási módjának elmeit, összevetem előnyeiket és hátrányaikat, és kitekintek a többdimenziós képességek tesztelésének adaptív lehetőségeire is.

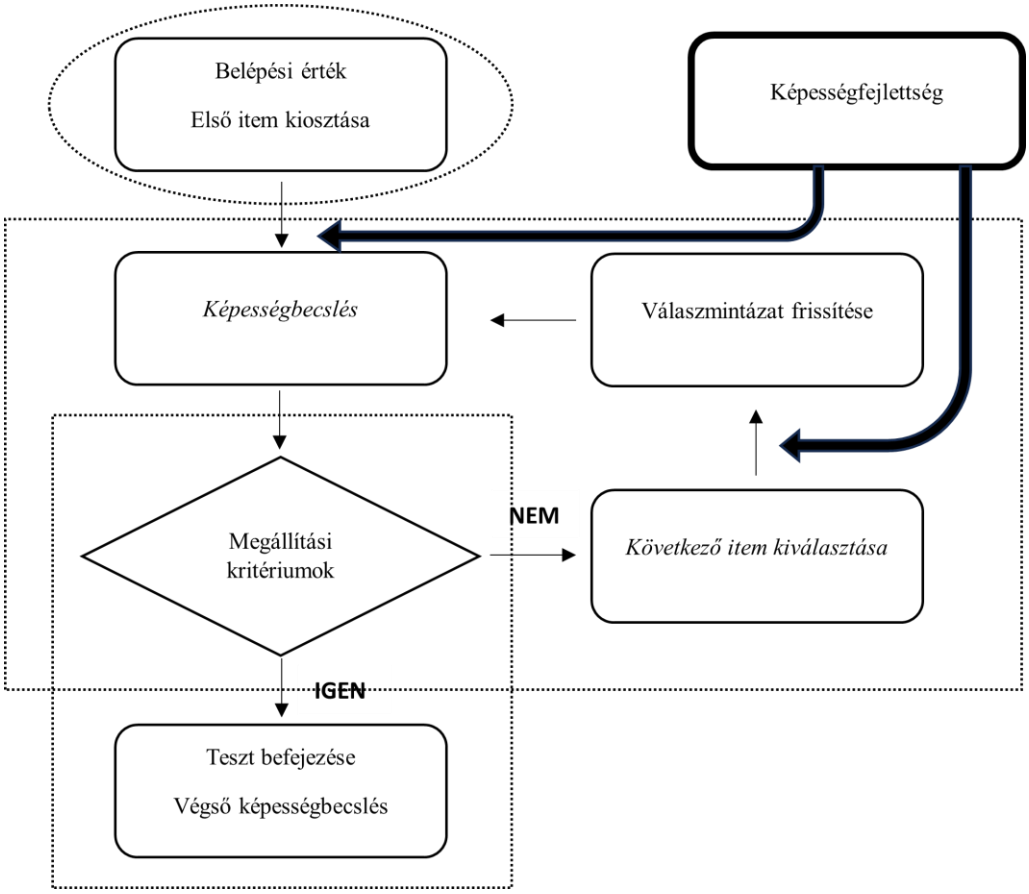
[3] Ennek a tulajdonságnak a TIMSS és PIRLS mérések csoportadaptív megvalósítása bemutatásakor lesz jelentős szerepe (ld. 3.2 fejezet).

2.3.1. A számítógépes adaptív tesztelés (CAT)

A számítógépes adaptív tesztelés (CAT) esetében a teszt egysége az item (Weiss & Kingsbury, 1984), az a legkisebb feladat, ami önállóan értékelhető. A számítógép az összes korábbi válasz alapján a válaszoló képességfejlettségét megbecsüli, majd a becsült értékhez valamilyen kritérium szerint leginkább illő következő itemet választja. A teszt ezen ciklusa addig tart, amíg valamilyen megállítási kritériumok, pl. adott számú item megoldása, a tesztre szánt maximális idő letelte vagy adott mérési pontosság elérése nem teljesül. A CAT tehát nem más, mint az adaptív tesztelés, az IRT módszerek és az interaktív számítógépes felmérésvezetés kombinációja. A CAT folyamatát a 4. ábra mutatja be.

4. ábra

A számítógépes adaptív tesztelés sematikus ábrája (Magis & Raîche, 2012) alapján



Megjegyzés. Balra felül a teszt induló része, középen a teszt ciklikus része a megállítási kritériumok teljesüléséig, alul pedig a teszt befejezésének szakasza. Vastaggal jelöltük a nem a rendszerből származó elemeket, vagyis a teszt kitöltőjének képességfejlettségét, amelyet becsülni szeretnénk, illetve ennek realizációját, az aktuális itemre adott választ.

A CAT-nak hat szerkezeti eleme van:

1) *IRT modell*, mely alapján az itemek paraméterei és a válasz alapján a válaszoló képességfejlettsége megbecsülhető (ld. 2.2 fejezet). A modellt mindig a teszt céljának megfelelően kell kiválasztani (Chang, 2015), ez tehát szakértői feladat. A választás eredménye a CAT esetében az itembank tartalmában (itemparaméterek), az itemek kiválasztásának folyamatában és a tanulói képességbecslés során jut szerephez. Az OKM a háromparaméteres IRT modellt alkalmazza.
2) *Itembank vagy feladatbank*, melyből az eljárás kérdései származnak. Az itemek jellemzőit a méréshez választott modellel kell meghatározni, nehézségüknek pedig a mérés céljához kell illeszkedniük. Kvalifikációs (sikeres/sikertelen) jellegű mérések esetében az itemek nehézségének elsősorban a határpont környékét, képességfejlettség mérés esetén a populációt jellemző teljes képességskálát fedniük kell, és célszerű, ha magas a diszkrimináló értékük (azaz a meredekségük). Az itembankkal kapcsolatos vizsgálatok egy része éppen a meredekség szükséges nagyságával foglalkozik. Általánosan néhány tíz item elegendő lehet, de 100 item már elegendő egy adaptív vizsgálathoz (Weiss & Kingsbury, 1984). 200–300 item már közepes nagyságú itembanknak mondható (Magyar, 2014b; Şahin & Weiss, 2015). Nagyobb vagy többször használt mérések esetén vizsgálható az itemek elhasználódása, mivel az itemek egy része sok tesztkitöltő előtt ismertté válik, ami a mérés biztonságát is veszélyeztetheti (Belov, 2014; Ozturk & Dogan, 2015). Ez a jelenség elsősorban a közepes nehézségű és nagy diszkrimináló képességű itemeket érinti.
3) *Kezdő vagy belépési érték*: az a képességpont, amelyhez az első item illeszkedik. A belépési érték lehet a teljes mintapopuláción ugyanaz, de ha rendelkezésre áll valamilyen előzetes információ, akkor ez is lehet személyre szabott. Az OKM esetében ilyen előzetes információ lehet a tanuló évfolyama, magasabb évfolyamok esetén a korábbi eredménye vagy az iskola típusa.
4) *Itemkiválasztási eljárás*: a becsült képességponthoz annak az itemnek a kiválasztása, amely megválaszolásától a képességbecslés legnagyobb javulása várható. A legáltalánosabb módszerek a maximum információ (Birnbaum, 1968; Weiss, 2011), a becsült képességponthoz legközelebbi nehézség szerinti választás (Urry, 1970), illetve a Bayesi megközelítés (van der Linden & Ren, 2019), melyek általában nagyon hasonló eredményre jutnak (Sympson et al., 1982 idézi Weiss &

Kingsbury, 1984; Thompson & Weiss, 2011). A itemkiválasztási eljárásokkal kapcsolatos vizsgálatok jellemzően különböző módszereket hasonlítanak össze a teszt befejezéséhez szükséges itemszám és/vagy a képességbecslés pontossága szerint, tipikusan szimulációs módszerekkel (Ito & Segall, 2013).

5) *Képességbecslési eljárás*: a rendelkezésre álló válaszok alapján a képességpont és esetleg a képességpont konfidencia-intervallumának becslése. A konfidencia-intervallum a becsült képességpontnak az a környezete, ahol a valódi képességpont bizonyos előre meghatározott valószínűséggel valóban megtalálható. A becslésre maximum likelihood és Bayes becslési módszereket használnak, esetleg ezek valamilyen kombinációját. A felmérés elején, kevés válasz esetén a Bayes módszer hatékonyabb, mivel nem érzékeny a szélsőséges válaszmintázatokra. A felmérés végén, több válasz esetén a maximum likelihood becslés pontosabb, mivel jobban figyelembe veszi a későbbi válaszokat (Magis et al., 2017a; Weiss & Kingsbury, 1984).
6) *Megállítási kritérium* (*stopping rule*) vagy *végződtetési kritérium* (*ending rule*) (Magyar, 2014b): az a feltétel, melynek teljesülésekor a teszt véget ér, a számítógép a felmérést befejezi. A teszt céljainak megfelelően több megállítási kritérium vagy ezek együttes teljesülése is lehet feltétel. Ilyen feltétel lehet, hogy a képességbecslés hibája bizonyos szint alá csökken (ez a feltétel minden válaszadóra egyformán pontos eredményt biztosít); a képességfejlettség konfidencia-intervalluma alapján a teszt kitöltője besorolható valamilyen képességfejlettség szerinti szintre (klasszifikáció); megtörtént bizonyos számú item megválaszolása; letelt a kérdések megválaszolására szánt maximális idő. Rövid tesztek és kisebb itembank esetén megállítási kritériumot indukálhat, hogy minden lehetséges item megválaszolásra került.

A CAT számos előnnyel kecsegtet. Elsősorban a lineáris tesztnél lényegesen rövidebb, akár fele olyan hosszú tesztkitöltéssel (Weiss, 2011) vagy ennek paraleljeként pontosabb képességbecsléssel. A kérdés nehézségének jobb megválasztása nagyobb motivációt eredményezhet, esetleg az érdeklődés fenntartását, ugyanakkor magasabb félelemmel is társulhat, amennyiben a teszt kitöltői nem rendelkeznek információval a CAT működési módjáról (Wise, 2014). Akhtar és munkatársai (2023) metaanalízisükben ilyen pozitívumot nem találtak.

A CAT módszernek előnyei mellett hátrányai is vannak (Magyar, 2012). Mivel az eljárás során minden item után értékelés történik, ami ráadásul számításigényes feladat, a CAT módszere olyan tesztek esetében alkalmazható, ahol az itemek szabadon választhatók, azaz nem köti össze őket közös kontextus. Adott szöveghez tartozó feladatok esetében ez az eljárás nem megfelelő, mert a megoldáshoz szükséges idő nagy részét a szöveg elolvasása teszi ki. Hasonló okból a CAT esetében nincs lehetősége a kitöltőnek visszalapozni, feladatokra visszatérni. Ez egyrészt ellehetetleníti azokat a tesztkitöltési stratégiákat, melyek az időigényesebb vagy nehezebb feladatok átugrására és későbbi megválaszolására épülnek, másrészt nagyobb stresszt okozhatnak a kitöltőben, ami rosszabb teljesítményhez vezethet, azonban Akhtar és munkatársai (2023) metaanalízisükben ilyen eltérést nem találtak.

Ezekre a problémákra alakult ki a többszakaszos adaptív tesztelés (*multistage adaptive testing*, MST) módszere. Az ausztrál NAPLAN mérés (ld. még 0 fejezet) többszakaszos adaptív módra történő cseréje előtt a közreműködő szervezetek motivációra vonatkozó vizsgálatokat végeztek. A kognitív interjúk alapján a tanulóknak nem okozott problémát a szakaszok eltérő nehézsége, valamint a visszalapozás korlátozott lehetősége, azonban erről a nehézségről a teszt megkezdése előtt tanácsos lehet tájékoztatni a tanulókat (Educational Assessment Australia (EAA), 2013). A 12 736 tanuló bevonásával történt vizsgálat nem talált különbséget a lineáris és adaptív teszten elért eredmény között, ugyanakkor az adaptív tesztút esetén a tanulók pozitívabb szubjektív élményről számoltak be, mint lineáris teszt esetén (Lifelong Achievement Group & Martin, 2015).

A CAT további nehézsége, hogy a papír-ceruza tesztekhez képest költséges a létrehozásuk (Magyar, 2012), elsősorban a számítógépes mérési rendszer és a nagyméretű, megfelelő itembank kialakítása miatt. A minél változatosabb és gazdagabb itembankok létrehozásának kérdése vezetett az automatikusan generált itemek, azaz egy eredeti item (szülő-item) automatikusan létrehozott változatainak (klón-itemek) vizsgálatához. Colvin és munkatársai (2016) szimulációs vizsgálatukban a szülő-item paramétereit használták a képességbecsléshez, azonban a tesztalanyok válaszait a klón-itemek kissé módosított paramétere alapján generálták. A szimulációk során mind a klón-itemek arányát, mind a szülő-itemétől eltérő viselkedés mértékét variálták, és eredményeik alapján sem a képességfejlettség és a becslés átlagos eltérése, sem a képességpont hibájának nagysága, sem a becslés hibája (*root mean square error*, RMSE) nem volt jelentős a csak szülő-itemeket használó tesztekhez képest. Ez alapján

lehetségesnek tartják a generált itemek használatát a szülő-item IRT jellemzőivel, méghozzá a paraméterek beméréséig (ha ez egyáltalán szükséges).

2.3.2. Többszakaszos adaptív tesztelés (MST)

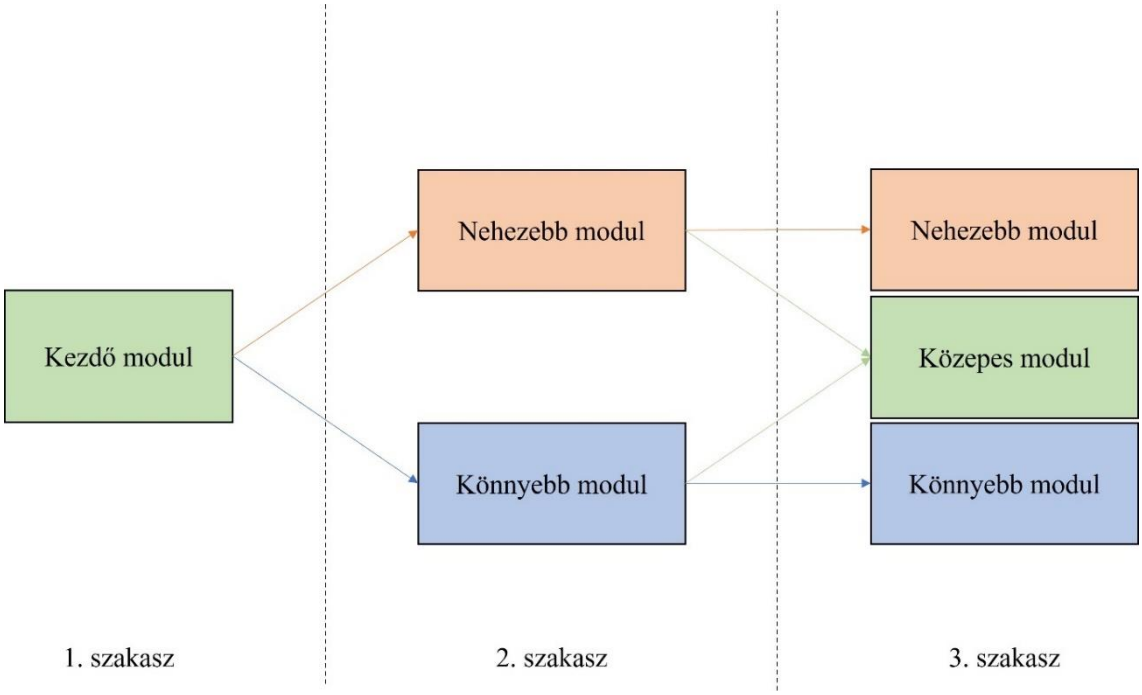
A CAT-tal szemben a *többszakaszos adaptív tesztelés* (MST) nem egy itemet, hanem néhány itemből álló modult adminisztrál a teszt kitöltőjének (Sari & Huggins-Manley, 2017; Yamamoto, Shin et al., 2018). További különbség, hogy a teszt nem ciklusba szervezett, hanem a tesztet előre meghatározott számú szakaszra bontják, a válaszok kiértékelése a szakaszok végén történik. A következő szint, azaz a teszt következő egységének nehézsége a teszt addigi eredménye alapján kerül kiválasztásra. Egy szakaszban a tesztalany egy egységet, azaz egy-egy előre összeállított itemcsoportot (Luecht & Nungester, 1998) old meg, melyet modulnak (pl. Han, 2020) vagy tesztletnek (pl. Frey et al., 2016) neveznek. A modulok nehézségük alapján szintekbe rendeződnek, azaz vannak könnyebb és nehezebb modulok. A többszakaszos adaptív tesztek legalább két szakaszból állnak és legalább két szintjük van (Magyar, 2012). A teszt akkor ér véget, amikor az előre kialakított tesztszerkezet, azaz adott számú szakasz véget ér.

Az MST mérés szerkezeti részei (5. ábra) a következők (Han, 2020):

1) *Blokk* (*block*): egy kontextushoz (azonos szöveghez vagy ábrához) tartozó feladatok.
2) *Modul* (*module*): egy egységbe szervezett itemek. Állhatnak független itemekből, blokkokból vagy egyetlen blokkból is. A modulok hossza lehet eltérő.
3) *Szint* (*level*): a mérés vertikális szervező elve, megfelel a CAT-ban az item nehézségének. Szakaszonként változó lehet a szintek száma, jellemzők a 2 és 3 szintű tesztek (Zenisky et al., 2010). Finomabb szintbeosztás is lehetséges, azonban szakaszonként legfeljebb négy szint jellemzően elegendő (Han, 2020).
4) *Szakasz* (*stage*): a teszt legnagyobb egysége, a mérés horizontális szervező elve. A szakasz lezárásakor történik a képességbecslés és a következő szint, illetve a szinten belül a modul kiválasztása. Az első szakaszt irányító (*routing*) szakasznak is szokás nevezni. Az MST teszt legalább két szakaszt tartalmaz, de gyakori a három és négy szakaszos szerkezet is (Han, 2020).

5. ábra

Egy tipikus MST teszt szerkezete. (Forrás: saját ábra)



Megjegyzés. 3 szakasz, ahol az első szakasz egy szintet, a második szakasz két szintet (könnyű és nehéz), a harmadik szakasz három szintet (könnyű-közepes-nehéz) tartalmaz. Az egyes szinteken kiosztott tesztrészek a modulok. A modulok itemeket vagy blokkokat (a modulnál kisebb itemcsoportokat) tartalmazhatnak.

A mérés részei nagyrészt megegyeznek a CAT-nál felsoroltakkal (1–5 jellemzők), de az itemek helyett modulok használata történik, ezért az egyes CAT elemek is módosulnak. A képességpont becslése az itemek helyett a modul végén történik, így adott szakaszon belül a kitöltő lapozhat a feladatok között. A következő item kiválasztását elsősorban a következő szint kiválasztása helyettesíti, ami történhet a teljes teszt és az utolsó teszt alapján is (Han, 2020), a modul kiválasztása ezután azonos minőségű modulok közül véletlenszerű, vagy, ha minden szakasz-szint kombinációhoz egyetlen modul van hozzárendelve, egyértelmű. Hasonló nehézségű itemek modulját alkalmazva a pontos képességbecslés helyett elegendő lehet a szakaszban jól megoldott itemek számát figyelni a következő szint kiválasztásának eljárásában (Gonthier et al., 2018; Yamamoto, Shin et al., 2018). A megállítási kritérium nem a képességpont becslésének függvénye: a teszt végét az utolsó szakasz lezárása vagy a tesztre fordítható idő vége jelenti.

A tesztkészítés eljárása is módosul a CAT-hoz képest. Első lépésben a teszt szerkezetét, azon belül a szakaszok és a szintek számát kell meghatározni. Ezután a szintek határait kell a képességskálán megállapítani. Az itembank kiegészül a modulok nyilvántartásával. Az egyes modulok összeállítása blokkokból vagy itemekből a teszt elindítása előtt történik úgy, hogy az egyes modulok a képességskála valamelyik szinthez tartozó intervallumát a lehető legegyenletesebben mérjék. Ehhez az egyes itemek paraméterei helyett a modul információs görbéjét (ld. 2.2 fejezet) használják fel (Yang & Reckase, 2020).

Az MST tesztekkel kapcsolatos kutatások jellemzően a szerkezettel (szintek és szakaszok száma, modulok hossza) és a teszt irányításával (következő szint választása) foglalkoznak. Mivel kevesebb beavatkozási pont van (a szakaszok végén), ezért a kezdő szakasz kialakítása kritikus. Vannak nagyobb téttel rendelkező MST technológiát használó mérések (pl. nyelvvizsgarendszerek) (Wang et al., 2019), ezért gyakoriak a tesztbiztonsággal foglalkozó vizsgálatok.

2.3.3. Továbblépési lehetőségek: többdimenziós adaptív tesztelés és válaszidő figyelembevétele

Az egydimenziós IRT mellett lehetőség van többdimenziós (MIRT) modellek alkalmazására az adaptív technológiák használata során (Frey & Seitz, 2009). A többdimenziós modellek azon alapulnak, hogy egy komplex területet mérő item nem egyetlen jellemzőt mér, hanem a megoldáshoz több alapvető készség együttes használata szükséges (Hartig & Höhler, 2009). Lehetséges, hogy adott item a különböző mért dimenziók szerint eltérő nehézségű, így az itemek jellemzőit annyi dimenzión kell nyilvántartani, ahány dimenziót mérünk. Ekkor mind az itembank itemeinek paramétereit, mind a képességbecslés módszerét, mind a teszt következő egységének kiválasztását a választott modellhez kell igazítani. MIRT modellek alkalmazhatók mind a CAT (Araci & Tan, 2022), mind az MST (Frey et al., 2016; Jewsbury & van Rijn, 2020) esetében. Frey és Seitz (2009) összegző írásukban úgy találták, hogy többdimenziós CAT esetében az itemek száma az adaptív teszthez képest további 30–50%-kal csökkenthető.

Számítógépes teszteléskor nincs akadálya a válaszadási sebesség mérésének, ami lehetőséget ad a kitöltő sebességének, mint képességnek vagy jellemzőnek a mérésére (Hornke, 2000). Ennek több felhasználása is lehetséges. Egyrészt a pontosság szempontjából megfelelő, de a teszt teljes időtartamát minimalizáló item kiválasztásában

(Finkelman et al., 2014; Sie et al., 2015), másrészt a motiválatlan, tippelő kitöltők azonosításában (Csányi & Molnár, 2021; Wise, 2014) is segíthet.

Bár a többdimenziós adaptív tesztek egyre inkább a mérésekkel összefüggő pszichológiai és neveléstudományi kutatások középpontjába kerülnek, jelen dolgozat keretei között nem foglalkozom velük részletesebben, mivel az OKM tartalmi kerete jelen állás szerint egydimenziós jelenségeknek tekinti a mért tantárgyi területeket. Ugyanakkor a jövőben érdemes lehet folytatni az ezirányú kutatásokat (pl. Kispál & Gergely, 2022), főleg, ha (mint az adaptív mérés esetében) fejlesztési cél a teszt további rövidítése, vagy egyszerre több mérési terület egyidejű tesztelése kevert műveltségi feladatsorok segítségével (Balázsi et al., 2021).

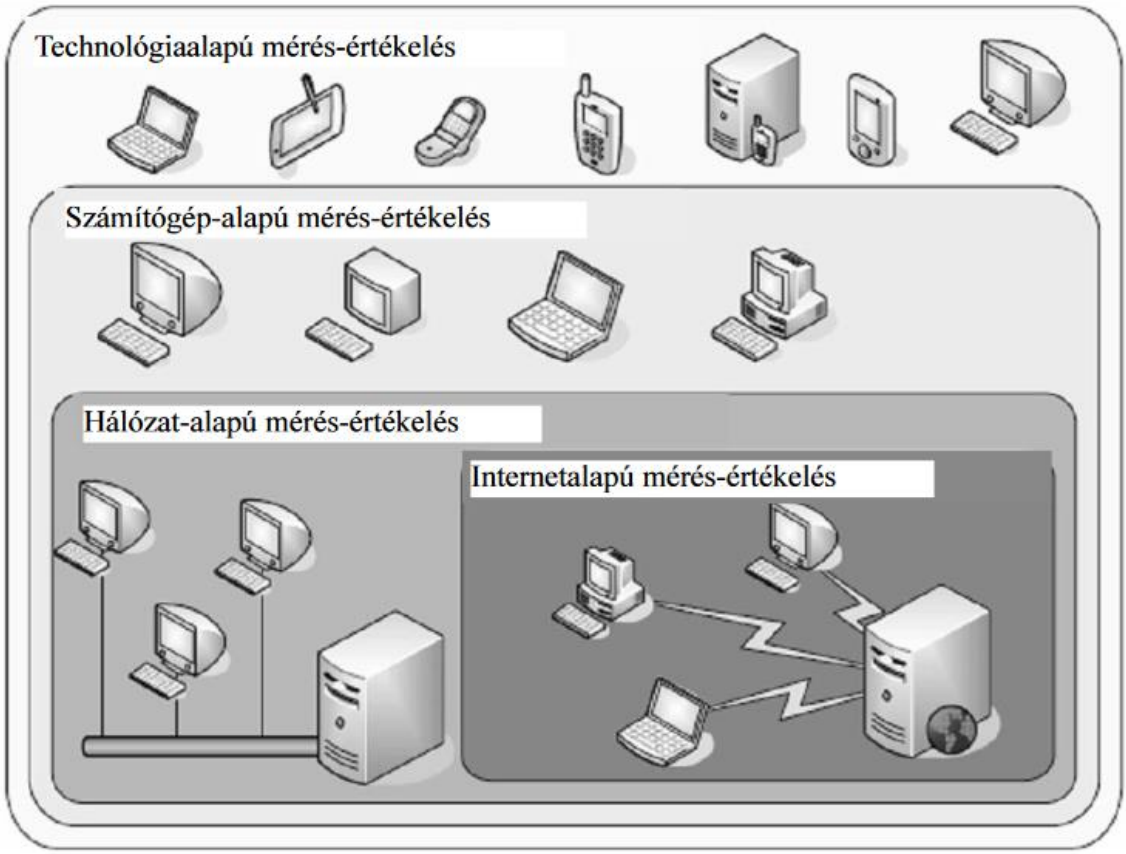
2.4. Technológia alapú, számítógépes, elektronikus vagy digitális mérés?

A technológia alapú, számítógépes, elektronikus, illetve digitális mérés kifejezések gyakran megjelennek a témával foglalkozó hazai szakirodalomban. A legbővebb fogalomnak a *technológia alapú tesztelés*t (*Technology Based Assessment*, TBA), illetve az *elektronikus tesztelés*t (*e-Testing*) tekinthetjük (Csapó et al., 2008). Ez minden olyan mérést magában foglal, ahol valamilyen technológiai/elektronikus eszközt használnak az adatgyűjtéshez, azonban ez a lehetséges eszközök széles palettáját jelenti a mobiltelefonoktól kezdve a szemmozgáskövető eszközökig. A legelterjedtebb eszköz a számítógép, azokat a méréseket, ahol a teszt felvétele ezen a felületen keresztül valósul meg, *számítógépes* vagy *számítógép alapú mérés*eknek (*Computer Based Assessment*, CBA) nevezzük. Ezen belül tovább finomítható a felosztás aszerint, hogy a mérés egyedi gépeken, egy kisebb, helyi hálózaton, vagy online, interneten keresztül valósul meg (6. ábra).

Az *elektronikus* jelző is gyakran megjelenik (pl. Molnár & Magyar, 2015), főként valamilyen szóösszetételben, mint elektronikus környezet, elektronikus tesztelés, elektronikus platform, melyek leginkább a „számítógépes” szinonimájaként használatosak, hozzátéve, hogy ezek a kifejezések általában nem önállóan definiálva jelennek meg, hanem a számítógép alapú tesztelés egyenértékű megfelelőjeként.

6. ábra

A technológiaalapú, a számítógépalapú, a hálózat- és internetalapú mérés-értékelés hierarchikus viszonya (Jurecka és Hartig, 2007 alapján) (Forrás: Csapó et al., 2008)



A *digitális* jelző leginkább tananyagokhoz, szövegekhez kapcsolódik, és azt fejezi ki, hogy a tartalom nem papír-ceruza vagy analóg adat. Szintén használatosak a digitális írástudás, digitális szövegértés vagy digitális műveltség kifejezések, amelyek a digitális információ értésére, alkotására vagy digitális eszközök használatára vonatkoznak. Ilyen értelemben a digitális adatok, tételek (itemek) a papír-ceruza tesztek adatainak, tételeinek (itemeinek) számítógépes tesztben használt változatai.

A disszertációban a *számítógépes mérés* kifejezést használom, mivel ez jól kifejezi az új közvetítő eszköznek a papír-ceruza teszttől való különbségét, illeszkedik a nemzetközi mérések és az OKM megvalósításához, és egyaránt vonatkozhat lineáris és adaptív tesztekre. Ahol nem a mérés közvetítőjére, hanem a tartalmára (itemekre, szövegekre) fókuszálok, ott a *digitális* jelzőt használom.

2.5. Médiahatás

A legáltalánosabb megfogalmazás szerint a *médiahatás (mode effect)* nem más, mint egy teszt kétféle adatfelvételi módja közötti pszichometriai különbség (Buerger et al., 2019). A két adatfelvételi mód jellemzően a papír-ceruza és a számítógépes adatfelvételt jelenti, de a technológiával támogatott mérések esetében az alkalmazott információ- és kommunikációtechnológiai (IKT) eszköz más (pl. tablet) is lehet (Hamhuis et al., 2020). A két mód közötti eltérésnek számtalan oka lehet: a szöveg megjelenésének különbségei (görgetés, szövegfülek) (Fishbein et al., 2018), a kézírás és a gépelés közötti különbség (Canz et al., 2020; Zehner et al., 2020), az itemek eltérő jellemzői (Buerger et al., 2019; Hülber & Molnár, 2013) vagy a megjelenítésre és bevitelre használt eszköz (Molnár et al., 2015). A médiahatás megjelenhet az egyes itemek szintjén (megoldottság, itemparaméterek, kihagyott feladat), a két tesztmédiumon elért teljesítmény közötti eltérésként (különbség, együttjárás mértéke) vagy a teszt jellemzőiben, a két módon mért konstruktum vagy a pszichometriai mutatók (pl. reliabilitás) eltérésében (Buerger et al., 2019). A médiahatás-vizsgálatok ezen különbségek feltárására és mérésére törekszenek.

Lineáris papír-ceruza tesztnek pontosan megfelelő lineáris számítógépes tesztre történő átalakítása esetén jellemzően nem változtatják meg a teszt további jellemzőit. Ennek megfelelően az egyes tesztrészek nagysága, a köztük tartott esetleges szünet vagy a korábbi tesztrészbe történő visszatérés hasonló módon történik. Lehetséges azonban, hogy a számítógépes megvalósítás lehetővé teszi a kihagyott vagy el nem ért itemek eltérő értékelését.

Papír-ceruza tesztek esetében egy itemről nem dönthető el teljes bizonyossággal, hogy a tanuló nem tudta megoldani, ezért üresen hagyta, vagy letelt a kitöltésre szánt idő, ezért nem is látta a feladatot. Ezeket az üresen hagyott feladatokat, hacsak az egész tesztrész nem üres, hibás megoldásnak tekintik, ezért a teljesítményt biztosan az alacsonyabb képességbecslés irányába mozdítják (Auxné Bánfi et al., 2014; OECD, 2014b). Számítógépes mérések esetében pontosan adminisztrálható, hogy a két lehetséges eset közül melyik történt, így lehetőség van arra, hogy az el nem ért itemek ne számítsanak bele a képességbecslésbe. Ez vezethet a két mérési mód közötti általános médiahatáshoz (Kroehne et al., 2019), ezért indokolt lehet az itemek esetében összehasonlítani a megválaszolatlan kérdések arányát (Fishbein et al., 2018)

3. A nagymintás tanulóiteljesítmény-mérésekről

A mérési és értékelési rendszerek kialakítása és fejlesztése nagy szakértelmet igényel, érvényességüket mind szakmai, mind módszertani szempontból magas tudományos színvonalon kell biztosítani. Az értékelésen belül a tanulóiteljesítmény-mérésének (*assessment*) különleges helye van (Halász, 2013). A többi lehetséges elérhető eszköz (pl. iskolai osztályzatok vagy munkaerőpiaci adatok) nem elég megbízható vagy nem kapcsolódik elég szorosan a konkrét iskola tevékenységéhez (Horn, 2010; Kertesi, 2008). Bár a teljesítménymérések jellemzően egy aspektus mérésére irányulnak, azonban megfelelő információkkal kiegészítve lehetőséget teremtenek az oktatás más fontos jellemzőinek, például a méltányosság vagy az esélyegyenlőség vizsgálatára (Fehérvári & Széll, 2014).

A mérés funkciójának térnyerésével a tanulóiteljesítmény-méréséhez tartozó standard tesztek fejlesztése és alkalmazása is robbanásszerűen fejlődött. A teljesítménymérések eredményét nem csak a döntéshozásban, de a tanfelügyeletek ellenőrző-minőségbiztosító funkciójának gyakorlása során is hasznosítják. Erre különösen a decentralizáltabb oktatási rendszerrel rendelkező országokban volt szükség az egységesebb oktatásminőség biztosítása érdekében.

A mérés-értékelés térnyerésével kialakultak az oktatási rendszerek összehasonlítását lehetővé tevő nemzetközi nagymintás tanulóiteljesítmény-mérések, és ezek mintájára nemzetállamok szintjén is megjelentek országspecifikus mérések. Az alábbi fejezetben elsősorban a papír-ceruza tesztektől a számítógépes, majd adaptív megvalósítás felé megtett fejlődés szemszögéből mutatom be azokat, melyekben Magyarország is érintett, illetve viszonyításai alapot jelenthetnek Magyarország számára.

3.1. Nemzetközi mérések digitalizációja

A Magyarország részvételével zajló nemzetközi nagymintás tanulóiteljesítmény-mérések a 2015. évi mérési ciklussal kezdődően nagy változáson mentek át a megvalósítás módját tekintve. A módszertani váltás során a papír-ceruza tesztekről számítógépes tesztekre, majd a PISA esetében adaptív tesztelésre tértek át. Az alfejezetben a változások sorrendjében röviden áttekintem a PISA, a PIRLS és a TIMSS mérésekben bekövetkezett változásokat. Mindhárom nemzetközi tanulóiteljesítmény-mérés hazai szervezője az Oktatási Hivatal.

3.1.1. PISA

A PISA az OECD által három évente megszervezett tanulóiteljesítmény-mérés. A legutóbbi mérési ciklus 2022-ben volt, az eredmények 2023 decembere óta elérhetők a szervezet honlapján[4]. A 2018-as mérés 79 országban vizsgálta a 15 éves tanulók szövegértési, matematikai és természettudományos műveltségét (OECD, 2019c). A PISA mérés esetében három mérési ciklust (2009, 2012, 2015) érdemes vizsgálni a számítógépes mérés bevezetése szempontjából.

2009-ben szerepelt először kiegészítő területként a digitális szövegértés (*digital reading*), amit weblap-szerű nemlineáris szövegekkel kapcsolatos feladatok segítségével mértek. A szövegeknek mind a formája, mind a szituációk szerinti megoszlása különbözött a két mérési módban (OECD, 2009). Az egyik ilyen különbség, hogy a digitális szövegértés esetében az iskolai célú szituációkhoz képest a nyilvános szituációk aránya nagyobb, mint a papír-ceruza tesztek esetében (OECD, 2009, 25–26). Hasonlóan, a gondolkodási művelet szempontjából a digitális szövegértés esetében a feladatok ötöde összetett feladat, míg a papír-ceruza változat esetében nincs ilyen típus, a különbözet leginkább az integrálás és értelmezés kategóriába tartozik (OECD, 2009, 43). A mérés hazai szervezője külön kötetet szentelt a digitális szövegértés területnek (Balázsi & Ostorics, 2011), melyben bemutatják a szövegértés és a digitális szövegértés területek mérése közti hasonlóságokat és különbségeket, valamint ismertetik a hazai és nemzetközi eredményeket. 2012-ben ismét választható mérési terület volt a digitális szövegértés, a skálát ekkor már a 2009-es digitális szövegértés skálához igazították a papír-ceruza mérés módszertanának megfelelően.

Bár a digitális szövegértés terület skáláit az összehasonlíthatóság kedvéért a nemzetközi átlag alapján a szövegértés skálához igazították, ezt nem előzte meg semmilyen médiahatás vizsgálat. Erre utal, hogy a mérés egyik dokumentumában sem jelenik meg a médiahatás angol megfelelője, a „*mode effect”* kifejezés. A digitális szövegértés területen alkalmazott itemek nem papír-ceruza itemek digitális változatai, ezért vizsgálatomban (ld. 6.1 fejezet) a két mérési területen elért eredmény összehasonlítását nem tekintettem médiahatás vizsgálatnak.

[4] https://www.oecd.org/pisa/publications/

A 2012-es ciklusban a hagyományos papír-ceruza mérések és a digitális szövegértés mellett bevezetésre került a matematika számítógépes mérése[5] (*computer-based assessment of mathematics*), mint választható terület (Balázsi et al., 2013; OECD, 2013a). A matematika fogalmát és tartalmi keretét kibővítették, hogy illeszkedjen hozzá az új médiumon mért konstruktum is. Az itemek megoszlása a műveleti, tartalmi és kontextuális kategóriák között azonos a két felmérésben. A matematika számítógépes mérése magában foglalja a számítógépes lehetőségek használatának képességét, azonban lehetőség szerint az IKT eszközhasználati igény elmarad a matematikai műveleteké mögött. Az itemek bizonyos része, de nem mindegyike, innovatív megoldásokat tartalmaz. A matematika számítógépes mérése skála kialakítása két lépésben történt. Először a papír-ceruza mérés pontszámai és transzformációja készült el, és ezzel párhuzamosan a számítógépes mérés pontszámai. Ezután a PISA szakemberei a papír-ceruza és a számítógépes mérés itemeit együtt is elemezték, ami a közös tartalmi keret miatt lehetséges. Az itemek átlagos nehézségét mindkét típusú mérésre kiszámították, és a két mérési típus közötti különbséggel eltolták a számítógépes mérés képességpontjait, majd alkalmazták a 2009–2012 ciklusok közötti transzformációt. Ezzel a számítógépes mérés képességpontjait a papír-ceruza méréshez, egyszersmind a matematika terület trendjéhez igazították, emiatt a két mérés eredményei összehasonlíthatók (OECD, 2014, 254). A matematika papír-ceruza és számítógépes mérése egységes, a külön skálák mellett kombinált skála is kialakításra került (OECD, 2014a, 40). A mérések hazai szervezője a digitális szövegértés és a matematika számítógépes mérése területeket a fő területekkel egy kötetben mutatja be, ismertetve a matematikai keret bővítését és az itemek készítésének szempontjait (Balázsi et al., 2013).

A 2015. évi PISA mérés teljes egészében számítógépes formában valósult meg (OECD, 2016b), országonként meghagyva a papír-ceruza mérés lehetőségét (OECD, 2017b). Minden mérési ciklus esetében az itemek egy része közös az előző mérési körben szerepeltekkel, és ugyanígy egy másik része bekerül a rákövetkező mérésbe. Ezeket a közös feladatokat trend vagy horgony itemeknek nevezzük, és ezek a feladatok szolgálnak arra, hogy az egyes mérési ciklusok eredményeit azonos skálára lehessen hozni, azaz az egyes mérési körök eredményei összehasonlíthatók legyenek. Azokban az

[5] A PISA mérések hazai összefoglalóiban következetesen „számítógépes matematika mérés” szerepel, ezt azonban nem tartom fogalmilag pontosan fedőnek, mivel a „számítógépes” jelző nem a műveltségi terület, hanem a mérés jelzője. Ennél fogva a későbbiekben a „matematika számítógépes mérése” változatot használom.

országokban, ahol a papír-ceruza mérési módot választották[6], a mérés kizárólag trend itemekkel valósult meg, vagyis a papír-ceruza tesztek kizárólag olyan itemeket tartalmaztak, amelyek a 2012. és a 2015. évi ciklusokban is szerepeltek. Ez a megoldás lehetővé tette a papír-ceruza mérést választó országok PISA 2015 eredményeinek trendekhez való illeszkedését és a többi ország eredményeivel való összehasonlítását. Az új, kifejezetten számítógépes mérési ciklushoz fejlesztett itemeknek, amelyek azután a 2015. és 2018. évi mérési körök közös itemei voltak, nem készítették el a papír-ceruza változatát (az innovatív itemformák esetében ez nem is feltétlenül lehetséges), így ezek a papír-ceruza tesztfüzetekben nem szerepeltek. Szintén ebben a mérési ciklusban vált először a háttérkérdőívek felvételének elsődleges módjává a számítógépes kitöltés.

3.1.2. PIRLS

Az IEA által ötévente megszervezett PIRLS mérés célja a 4. évfolyamos tanulók szövegértési képességének felmérése, melynek 2021-es mérési ciklusában 65 ország és oktatási rendszer vett részt (Mullis et al., 2023). 2016 óta a mérésnek része a PIRLS Literacy (Mullis & Martin, 2015), mely tartalmi keretében megegyezik a szövegértés méréssel, de könnyebb kérdéseket tartalmaz, így a teljesítményskála alsóbb részére irányul. Ez a mérés olyan országokban választható, ahol 4. évfolyamon még az alapvető olvasási képességek fejlesztése zajlik. Az egyező (horgony) feladatok miatt a PIRLS és a PIRLS Literacy skálái összeköthetők, az egyes országok eredményei összehasonlíthatók.

A digitalizáció szempontjából a 2016-os és 2021-es mérések a mérvadóak, mivel 2016-ban került bevezetésre az ePIRLS kiegészítő mérés 14 ország részvételével (Mullis & Martin, 2015), 2021-ben pedig először volt elsődleges a számítógépes adatfelvétel (Mullis & Martin, 2019). A mérés dokumentumaiból kiderül, hogy a választható ePIRLS mérés célja kissé eltér a papíralapú mérésétől, mivel az iskolai környezetben folytatott, interneten történő olvasás során megnyilvánuló szövegértést vizsgálja (Mullis & Martin, 2015, p.22). Ehhez igazodva mind a teszt felépítése, mind az összetétele különbözik a főmérésétől. Egyrészt, a lineáris, nyomtatott szövegek helyett internetszerű, weboldalakat szimuláló felületen böngésznek a tanulók. Ennek következtében a mérés tartalmaz – a szövegértés mellett, vagy annak részeként – olyan stratégiai elemeket, mint egy még meg nem tekintett oldal információjának megbecslése, vagy a korábban megtekintett oldalakon talált információkra való emlékezés. Másrészt, a mérés céljához igazodva, a

[6] 2015-ben 72 ország és oktatási rendszer közül 15 élt a papír-ceruza mérés lehetőségével.

szövegek teljes egészében információkeresési célúak, eltérően a főméréstől, ahol a szövegek egyik felének célja az élményszerzés, a másik felének pedig az információszerzés (Mullis & Martin, 2015, Exhibit 2.). A terület eredményeit az itemek szintjének meghatározásával a szokásos PIRLS skálához igazítják (Martin et al., 2017, 13. fejezet), ezáltal a két terület eredményei összehasonlíthatók. A szervezők nem végeztek médiahatás-vizsgálatot a két mérés összehasonlításával. A két terület eltérő tartalma és mérése alapján a PIRLS és az ePIRLS méréseket nem tekintem ugyanazon mérés papír-ceruza és számítógépes változatának (ld. 6.1 fejezet), így álláspontom szerint az eredmények összevetését nem lehet médiahatás-vizsgálatnak tekinteni. A 2016-os ePIRLS mérésben Magyarország nem vett részt, így a méréshez kapcsolódó hazai jelentésben az ePIRLS-t meg sem említik (Balázsi et al., 2017).

A PIRLS 2021 mérési ciklusban az adatfelvétel a 65 résztvevő ország közül 33-ban (köztük Magyarországon is) számítógépes (digitalPIRLS), 32-ben papír-ceruza (paperPIRLS) formában valósult meg (Mullis et al., 2023). Az összehasonlíthatóság érdekében a két mérés tartalmában megegyezik. A digitalPIRLS és ePIRLS területek különbségét erősíti meg, hogy az ePIRLS továbbra is külön mérési területként szerepel, azonban a szervezők kifejezik a terület integrálása iránti szándékukat (Mullis & Martin, 2019, 67) PIRLS és ePIRLS hibrid tesztfüzetekkel.

3.1.3. TIMSS

Szintén az IEA szervezésében négyévente kerül sor a TIMSS vizsgálatra, mely 4. és 8. évfolyamon méri a tanulók matematikai és természettudományi teljesítményét (Mullis & Martin, 2017). A 2019. évi mérésben 64 ország vett részt (Mullis et al., 2020)[7]. A TIMSS mérésnek 2015 óta része a könnyebb feladatokat tartalmazó TIMSS Numeracy, mely 4. évfolyamon méri a matematika területet olyan országokban, ahol ezen az évfolyamon még az alapvető számolási készségek fejlesztése zajlik. A két mérés eredményeit, hasonlóan a PIRLS és PIRLS Literacy méréshez, azonos teljesítmény-skálán lehet megjeleníteni.

A digitalizáció szempontjából meghatározó a 2019. évi mérési ciklus, mivel e mérés során vezették be a számítógépes mérési rendszert. Ugyanebben a mérési rendszerben kerül megszervezésre 2021-től a PIRLS mérés is. A TIMSS mérés 2019-ben

[7] A 2023. évi mérési körben 72 ország vett részt. Az eredmények 2024 decemberében lesznek nyilvánosak (https://www.iea.nl/studies/iea/timss/timss2023).

magában foglalta az elsődleges számítógépes mérést (eTIMSS), amiben az országok több, mint fele vett részt, a papír-ceruza mérési módot (paperTIMSS) és a TIMSS Numeracy mérést.

3.1.4. Összegző tapasztalatok

A fentiekben említett három nemzetközi nagymintás tanulóiteljesítmény-mérés az utóbbi mérési ciklusok során a papír-ceruza tesztfüzeteket számítógépes tesztekre cserélte. A különböző mérések esetében eltérő indoklással találkozhatunk. A munka világában történő helytálláshoz szükséges kompetenciák fejlettségének mérését célul tűző PISA esetében bizonyos 21. századi képességek mérésére (mint az internetes szövegek értése vagy az interaktív matematikai eszközök használata) alkalmasabbnak találták a számítógépes megvalósítást (OECD, 2017a). A PIRLS esetében a különböző nehézségű mérések (PIRLS és PIRLS Literacy) és a 2016-ban bevezetett ePIRLS egységes szervezési és megvalósítási lehetősége indokolta a váltást (Mullis & Martin, 2019). A TIMSS esetében a természettudomány terület mérését kiegészítő innovatív, kísérleti szituációkat szimuláló problémamegoldás terület bevezetése, valamint a különböző nehézségű mérések egységes keretben való szervezése volt a fő indok (Mullis & Martin, 2017).

A vizsgált nemzetközi mérések tapasztalatai alapján a számítógépes mérésre átállás folyamatának legfontosabb vizsgálati kérdései, hogy: (1) Azonos maradt-e a mérés által vizsgált konstruktum? (2) Alkalmas maradt-e a mérés trendvizsgálatokra, azaz hogyan lehet összekötni az új mérési felületről származó eredményeket a korábbiakkal? (3) Eltérő-e az átállás hatása a különböző alpopulációkban – nemzetközi mérések esetében az egyes országokban – vagy az egyes mérési területeken?

A fenti kérdések vizsgálata más mérési rendszerek esetében is releváns, a nemzetközi mérések eredményeinek ismerete és a módszertan jó gyakorlatként felhasználható. A kérdésekhez kapcsolódó vizsgálat eredményét a 6.1 fejezet tartalmazza. A papír-ceruza tesztek számítógépes tesztekre történő cseréje lehetővé teszi egyrészt az innovatív itemek (drag-n-drop, legördülő menü, grafikus és multimédiás elemek) használatát, másrészt az adaptív tesztek bevezetését.

3.2. Adaptív mérési rendszerek bevezetése a nemzetközi mérések esetében

A számítógépes mérések a TIMSS és PIRLS esetében a csoport adaptív (*group adaptive*) (Mullis et al., 2021; Mullis & Martin, 2019), a PISA esetében a többszakaszos adaptív mérések (*multistage adaptive testing*) irányába mozdultak el (OECD, 2019a).

A csoport adaptív mérés a különböző szintű tesztelemek (itemek, feladatlapok, tesztfüzetváltozatok) vagy különböző célú (pl. TIMSS és TIMSS Numeracy) mérések egységes rendszerben történő fejlesztését, szerkesztését teszi lehetővé. A mérés ugyanebben az egységes rendszerben valósul meg, azaz a tanuló országának (vagy valamely egyéb jellemzőjének) ismeretében a rendszer a megfelelő, a könnyebb feladatokat tartalmazó TIMSS Numeracy tesztfüzet kitöltését teszi valószínűvé. A TIMSS és PIRLS csoport adaptív mérése nem tekinthető a 2.3 fejezetben definiált adaptív tesztelésnek, mivel maguk a kitöltések hagyományos lineáris tesztváltozatok, nem a mérés során a becsült képesség alapján dinamikusan előálló tesztek. A mérési dokumentumokban nincs arra utaló jel, hogy a jövőben a csoport adaptív tesztelésről az adaptív tesztelés irányába további fejlesztés várható, ugyanakkor a PIRLS 2021 eredményei azt mutatják, hogy a csoport adaptív megvalósítás a PIRLS 2016 nem adaptív megvalósításához képest a teljesítmény pontosabb becslését és a kihagyott feladatok kisebb arányát eredményezte (von Davier et al., 2023).

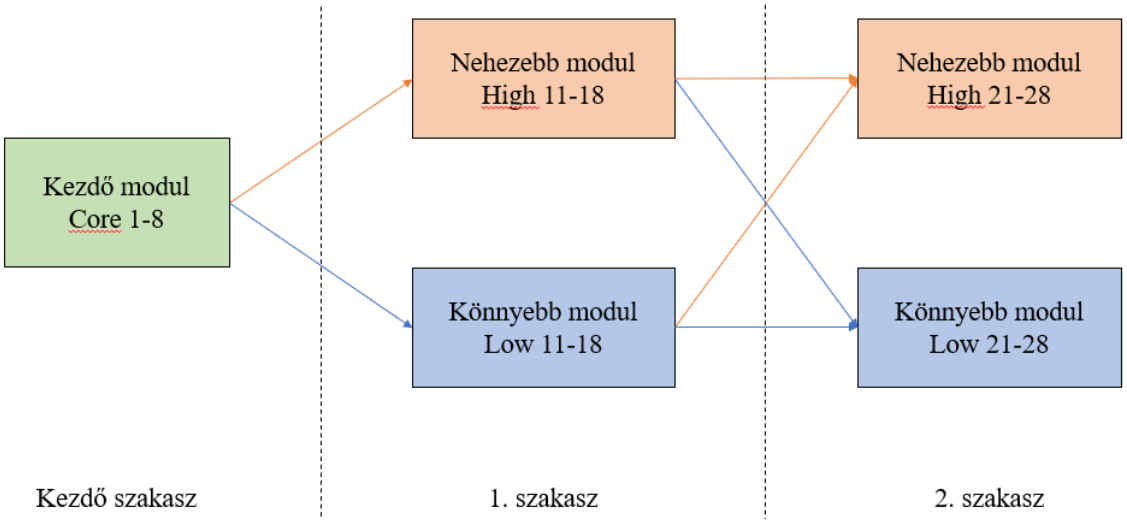
A PISA 2015 számítógépes tesztelésre történő átállása utáni következő lépés, hogy a lineáris tesztek helyett a fő mérési területen (2018-ban a szövegértés) többszakaszos adaptív tesztelést (MST[8]) alkalmaztak (OECD, 2019d; Yamamoto, Shin et al., 2018). A számítógépes adaptív (CAT) és a többszakaszos adaptív (MST) módszer közötti választást – az MST javára – az indokolta, hogy ez a tesztszerkezet jobban illeszkedik a PISA feladatszerkesztéséhez, mivel az MST itemek előre összeállított csoportját (modul) választja egyetlen item helyett, így egy egységet képeznek az egy szöveghez tartozó kérdések vagy egy egységbe szervezett matematika itemek. A másik indok, hogy ezt az eljárást az OECD felnőttek szövegértés és matematika műveltségét vizsgáló Programme for the International Assessment of Adult Competencies (PIAAC) mérésében már vizsgálták és jó eredménnyel kipróbálták (Yamamoto, Shin et al., 2018).

[8] A PISA publikációk a többszakaszos adaptív tesztelésre az „MSAT” rövidítést alkalmazzák, azonban az egységes nyelvezet érdekében az „MST” jelölést használjuk. A folyóiratokban megjelenő publikációkban maguk a szerzők is ezt a jelölést alkalmazzák (Yamamoto, Khorramdel et al., 2018; Yamamoto, Shin et al., 2018).

A PISA 2018 szövegértés területének MST mérése a teszt szerkezetének szempontjából 1–2–2 felépítésű, azaz a kezdő szakasz utáni két tesztszakaszban egy-egy könnyebb és nehezebb szint kapott helyet (7. ábra). Minden szakaszhoz és azon belül minden szinthez 8–8 lehetséges modul[9] tartozik, a modulok pedig kisebb egységekből épülnek fel. A modulok között átfedés van, ami az egyes tesztváltozatok közös skálán történő elhelyezéséhez szükséges.

7. ábra

A PISA 2018 szövegértés terület többszakaszos adaptív tesztjének szerkezeti ábrája (A verzió). (Forrás: saját ábra (OECD, 2019d) alapján)



A többszakaszos adaptív tesztelés minden résztvevőhöz véletlenszerűen választ egyet az azonos nehézségű kezdőmodulok (Core 1 – Core 8) közül. A modulok 2 egységből, az ötféle egység mindegyike 3–5 itemből áll össze. Egy modulban 7–9 automatikus kódolású item kapott helyet. A kezdő (Core) szakasz kitöltésének sikeressége és a következő szakasz szintje az automatikus kódolású itemek válaszai alapján került megállapításra. Adott kezdő modul után csak bizonyos modulok következhetnek.

Az első szakaszban 8 nehezebb és 8 könnyebb modul van. Az első szakasz moduljaiban a modul 3 egységből, a 24 féle egység mindegyike 3–6 itemből áll össze. Egy modulban így összesen 8–11 automatikus kódolású item kap helyet, és adott egység

[9] A PISA technikai leírása (OECD, 2019d) a „testlet” kifejezést használja, azonban az egységes nyelvezet érdekében a 2.3.2 fejezetben bemutatott „modul” jelölést használjuk.

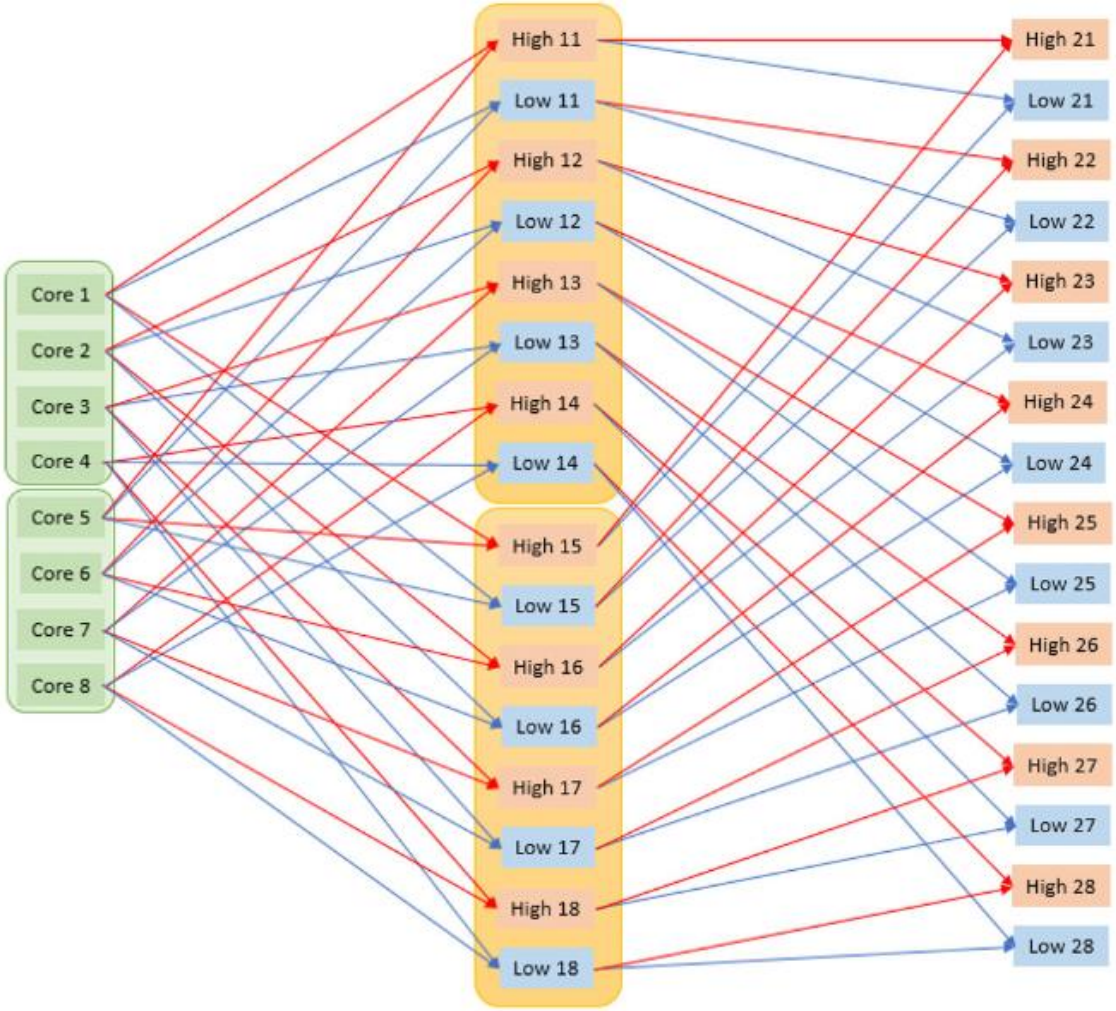
csak nehezebb vagy könnyebb modulokban szerepel. A második szakasz szintje a kezdő (Core) és az első szakasz kitöltésének sikeressége, azaz a két szakaszban az összes automatikus kódolású item jó válaszainak száma alapján kerül megállapításra. Adott első szakaszbeli modul után ismét csak bizonyos modulok következhetnek.

A második szakaszban ismét 8 nehezebb és 8 könnyebb modul szerepel. Egy modul 2 egységből, a 16 féle egység mindegyike 5–8 (jellemzően 7) itemből áll össze. Egy modulban 6–12 automatikus kódolású item van, és az egységek fele nehezebb és könnyebb modulban is szerepel.

Példa: a Core 1 modulban 9 automatikus kódolású item van. Legfeljebb 3 jó válasz esetén a könnyebb, Low 11 vagy Low 15 könnyű modul, legalább 7 jó válasz esetén a nehezebb, High 11 vagy High 15 nehéz modul következik az első szakaszban. A 4–6 jó választ adók esetében a következő szint és modul véletlenszerűen kerül kisorsolásra. Tegyük fel, hogy a Core 1 modul után a tanuló a High 11 modult tölti ki. A Core 1 és High 11 modulban összesen 20 automatikus kódolású item van. Legfeljebb 7 jó válasz esetén a könnyebb Low 21 modul, legalább 14 jó válasz esetén a nehezebb High 21 modul következik a második szakaszban. A 8–13 jó választ adók esetében a következő szint véletlenszerűen kerül kisorsolásra. A modulok pontos egymásra épülését a 8. ábra mutatja be. Ezen azt is megfigyelhetjük, hogy a Core 1 kezdőmodullal induló tesztek mindegyike a High 21, High 22, Low 21 és Low 22 modulokkal záródik, és az első szakasz négy lehetséges moduljával összesen 8 tesztváltozat tartozik hozzá. Ezek a tesztváltozatok megegyeznek Core 5 kezdőmodulhoz tartozó tesztváltozatokkal, és a teljes tesztet tekintve összesen 64 különböző tesztverzió lehetséges.

8. ábra

A PISA 2018 szövegértés terület többszakaszos adaptív teszt moduljainak kapcsolati ábrája (A verzió). (Forrás: OECD, 2019e)



Az egyes szintekhez szükséges helyes válaszok száma a modul karakterisztikus görbéje és a modulban szereplő automatikus kódolású itemek száma alapján került meghatározásra, a vágópontok a 425 és 530 képességponthoz tartozó várható eredmények. Az alacsony és magas sikeresség esetén sem volt egyértelmű a következő szint meghatározása, az esetek 10 százalékában a jó válaszok száma alapján egyértelműen meghatározott szint helyett a másik szint moduljával folytatódott a tesztkitöltés. Ez annak érdekében történt, hogy a nagyon eltérő teljesítményű országok tanulói által kitöltött tesztváltozatok között is biztosan legyen elegendő átfedés, vagyis az országok továbbra is ugyanazon a képességskálán legyenek megjeleníthetők.

Az itemek pozíciójával kapcsolatos vizsgálatok érdekében kidolgozásra került a modulok egy másik kapcsolati ábrája (B verzió). Ebben az első és a második szakasz

moduljai helyet cserélnek, de összességében ugyanazok a tesztváltozatok állnak elő. A B verziót a kitöltők 25%-ának osztották ki.

3.3. További nagymintás vagy tétttel rendelkező adaptív mérések

A nemzetközi nagymintás tanulóiteljesítmény-mérések célja országok vagy oktatási rendszerek összehasonlítása egymással vagy korábbi eredménnyel, ennél fogva a tesztek kitöltői, a tanulók szempontjából alacsony, úgyszólván semmilyen tétttel nem rendelkeznek. Hasonló a helyzet a magyarországi OKM esetében is, mely mérési egysége az iskola. Bár korábban voltak olyan példák, ahol a tesztfüzeteket a tanárok lepontozták és az összpontszám alapján osztályzatot adtak, azonban ezt a gyakorlatot az Oktatási Hivatal kerülendőnek minősítette, részben a képességbecslés nagy egyéni hibái miatt.

Az amerikai egyesült államokbeli National Assessment of Educational Progress (NAEP)[10] mérés kétévente nemzeti és állami szinten vizsgálja a 4. és 8. évfolyamosok körében az olvasás, az írás és a matematika területeket, valamint trendvizsgálatokat végez a 9, 13 és 17 évesek körében. A tanulókat három teljesítményszintre sorolják be, az eredményeket a The Nation's Report Card visszajelző rendszeren keresztül a tanulók és szüleik is megismerhetik, az eredmények a NCLB intézkedéseinek indikátorai voltak (Tomasz, 2011). A NAEP 2017-ben tért át a számítógépes mérésre az olvasás és a matematika területeken (The National Assessment Governing Board, 2017), azonban ennek menetéről a honlapon vagy adatbázisokban nem találtam leírást. A NAEP esetében nem történt meg az adaptív mérésre történő áttérés, azonban egy kétszakaszos három szintű (1–3) mérési szerkezetet 1993-ban kipróbáltak (Bock & Zimowski, 2003), aminek bevezetését javasolták a képességskála alacsonyabb részének pontosabb mérésére (Daro et al., 2007). Állami szinten léteznek olyan tesztek, melyek megvalósítása adaptív (Kingsbury & Hauser, 2004).

Vannak olyan, a kitöltő szempontjából nagy tétttel rendelkező tesztek, melyek adaptív mérés formájában valósulnak meg. Ide sorolható a NAEP-hez hasonló NAPLAN[11], mely 2008 óta évente méri Ausztráliában a 3., 5., 7. és 9. évfolyamosokat olvasás, számolás, helyesírás és nyelvtan/központozás területeken. A mérés 2018-ban vezette be először a számítógépes mérést, párhuzamosan az adaptív teszteléssel (ACARA, 2020). 2019-ben a tanulók fele töltött ki számítógépes tesztet. A próbamérések

[10] https://www.nagb.gov/naep
[11] https://www.nap.edu.au/naplan

után egy lényegében három szakaszos három szintű (1–2–3) szerkezet mellett döntöttek, néhány területen kiegészítve egy negyedik, legalsó szinttel átjárás nélkül a nehezebb tesztutak felé (ACARA, 2014). A tanulókat négy szintre sorolják, és a teszt eredménye hatással lehet arra, hogy a tanuló felsőbb osztályba léphet-e.

Még nagyobb tétttel bír a Test of English as a Foreign Language (TOEFL) nyelvvizsga, mely szintén többszakaszos adaptív mérési móddal zajlik (ETS, 2023). Az angol nyelv ismeretét és használatát méri hallott szöveg értése, olvasás, írás és beszéd területeken, az eredményét számos egyetem elfogadja a külföldi jelentkezők nyelvismeretének igazolására. Az adaptív szerkezet első vizsgálata már 1989-ben megtörtént (Hicks, 1989). A háromszakaszos ötszintes szerkezetet (1–2–4) két tesztrészen (írás és szerkezet, szókincs és olvasás) kontrollcsoportos vizsgálatban próbálták ki. A másik két tesztrészt a hagyományos módon mérték. A teszt fele annyi itemet használt, mint a papír-ceruza változat, ennek ellenére a papír-ceruza tesztekhez hasonló eredmények születtek, a képességskála szélein pedig pontosabbak voltak az eredmények. Bár ekkor még a számítógépes kapacitás miatt mint ritka alternatívát ajánlották (Hicks, 1989), később ez vált általános eljárássá, sőt az egyik első Japánban bevezetett adaptív teszttté (Nogami & Hayashi, 2010). Ennek hatására más méréseket átalakítottak vagy eleve adaptív teszteléssel terveztek, mint például a Computerized Assessment System for English Communication (CASEC) angol nyelvi mérést, mely CAT szerkezettel valósult meg. A teszt szakaszai 30 helyett átlagosan 22 itemmel és nagyobb mérési pontossággal futottak le (Nogami & Hayashi, 2010).

3.4. Magyarországi tanulói mérések

Magyarországon többféle tanulói mérés működik. Bander és munkatársai (2015) az országos mérések rendszere részének tekintették az Országos kompetenciamérést, az érettségi rendszerét és a Diagnosztikus fejlődésvizsgáló rendszert (Difer). Mára ide sorolható a Nemzeti Egységes Tanulói Fittségi Teszt (NETFIT), az idegen nyelvi és a célnyelvi mérés, valamint a Szegedi Tudományegyetemen fejlesztett elektronikus Diagnosztikus mérési rendszer (eDia) is. 2022-től kezdődően az Országos kompetenciamérés, valamint az idegen nyelvi és a célnyelvi mérés egységes digitális mérési rendszerben, egyre bővülő területekkel és évfolyamokkal, teljes egészében számítógépes mérésként kerül megszervezésre[12].

[12] https://www.oktatas.hu/kozneveles/meresek/digitalis_orszagos_meresek/altalanos_leiras

3.4.1. Difer, NETFIT, idegen- és célnyelvi mérések

A Difer mérés[13] az 1. évfolyamos tanulók körében az iskolai előrehaladás szempontjából kritikus elemi készségek felmérésére szolgál. A mérés nem teljes körű, a pedagógusok évente (jellemzően novemberben) körülbelül a tanulók egyharmadát mérik fel.

A NETFIT[14] 2015 óta méri teljeskörűen az 5–12. évfolyamokon a diákok fizikai állapotát és edzettségét 4 profilban. A mérés az Oktatási Hivatal és a Magyar Diáksport Szövetség együttműködésével valósul meg. A felméréseket a pedagógusok végzik minden év január és április között.

Az idegennyelvi mérés[15] és a célnyelvi mérés[16] 2015 óta az általános iskolák 6. és 8. évfolyamán teljes körűen lebonyolításra kerülő nyelvi mérés (Balogh, Garay-Madarász et al., 2021). A mérés célja feltérképezni a tanulók nyelvi készségeit az első idegen nyelv (angol vagy német) vagy a tanításban használt idegen nyelv (angol, német vagy kínai) területén. A kínai nyelv tartalmi kerete többször változott, elsősorban az íráskészség mérésének és fontosságának aránya. A nyelvi készség elvárt szintje az első idegen nyelv esetén a 6. évfolyamosok körében a Közös Európai Referenciakeret[17] (továbbiakban: KER) szerinti A1 szint, a 8. évfolyamon a KER szerinti A2 szint, a tanításban használt idegen nyelv esetén a 6. évfolyamosok körében a KER szerinti A2 szint, a 8. évfolyamon a KER szerinti B1 szint. Mért évfolyamai 2023-ban a 7. évfolyammal egészült ki, 2024-től kibővülnek a 6–11. évfolyamokra, ezzel összhangban az elvárt KER szintek is kiegészítésre kerültek.

Az idegennyelvi és célnyelvi méréseket 2021-ig a pedagógusok vették fel, majd az iskolák megküldték az eredményeket az Oktatási Hivatalnak. 2022-től a digitális országos mérések felületén kerülnek lebonyolításra és az eredményeket az OKM-hez hasonló mérésmódszertannal készítik és ismertetik (Oktatási Hivatal, 2023a). Az eredményekről az iskola júniusban visszajelzést kap.

A fenti mérésekben és az OKM-ben érintett évfolyamokat és mérési területeket, valamint a mérések időpontját a tanév rendjéről szóló rendelet (pl. 30/2023. (VIII. 22.) BM rendelet - a 2023/2024. tanév rendjéről, 2023) szabályozza. Mindegyik mérést az Oktatási Hivatal szervezi, és a pedagógusok folytatják le. A mérés eredményeiről jelentés

[13] https://www.oktatas.hu/kozneveles/meresek/difer/difer_leiras
[14] https://www.netfit.eu/public/pb_about.php
[15] https://www.oktatas.hu/kozneveles/meresek/idegen_nyelvi_meres/tartalmi_keret
[16] https://www.oktatas.hu/kozneveles/meresek/celnyelvi_meres/tartalmi_keret
[17] https://nyak.oh.gov.hu/nyat/doc/ker_2002.asp

készül, melyet az Oktatási Hivatal saját honlapján, az adott mérésnél, a NETFIT esetében annak saját felületén vagy (2021-ig) az OKM eredményeit tartalmazó felületen tesz közzé[18].

3.4.2. Elektronikus Diagnosztikus mérési rendszer (eDia)

Az elektronikus Diagnosztikus mérési rendszer[19] (Dia, később eDia) a Szegedi Tudományegyetem Oktatáselméleti Kutatócsoportjának fejlesztése (Molnár & Csapó, 2019), egy online segítő-fejlesztő diagnosztikus rendszer, amely nemcsak egy mérésértékelési platform, hanem egy tartalommal feltöltött rendszer (Molnár, 2015). Az eDia az iskolák önkéntes részvételével vizsgálja az 1–6. évfolyamokat elsősorban a matematika, a szövegértés és a természettudomány területén, azonban az egymást követő projektek során a mért területek bővültek és maga a rendszer is fejlődött. Az eDia más mérések, például a szegedi longitudinális mérések platformjául is szolgál. E disszertációban nem célom az eDia tartalmi keretének vagy céljának vizsgálata, ahogy a nemzetközi tanulóiteljesítmény-mérések és az OKM esetében sem. Ebből kifolyólag a következő alfejezetekben az eDIA mérési rendszer fejlesztése során végrehajtott kutatásokra, azaz a papír-ceruza és számítógépes mérés közti különbségek vizsgálatára, és az adaptív tesztelés lehetőségeinek feltárására irányuló elemzésekre fókuszálok.

3.4.3. Az eDia projektek digitalizációhoz kapcsolódó vizsgálatai

Magyarországon elsősorban az eDia mérés-értékelési keretrendszerben megvalósuló mérésekhez kapcsolódnak médiahatás vizsgálatok. R. Tóth és Hódi (2011) – akkor még a TAO-platformokon (Testing Assisté par Ordinateur) – 6. évfolyamos tanulók esetében szignifikáns, átlagosan csaknem 5%-os eltérést talált a két módban felvett szövegértés tesztek százalékos megoldottsága között. Ez körülbelül a szórások 44%-ának felelt meg, tehát a különbség hatásmértéke közepesnek mondható. Ez alapján a diákok átlagosan jobb eredményt értek el a nyomtatott médiumon, mint számítógépen. Eredményeik szerint a nem folyamatos szövegek esetén, a diagramleolvasási feladatoknál a legjelentősebb különbség, a legkisebb pedig a táblázatba foglalt információk megértésénél volt tapasztalható.

[18] https://okm.kir.hu/fit/
[19] https://edia.hu/?q=hu/elektronikus_diagnosztikus_meresi_rendszer

Hülber (2012) 1–6. évfolyamos tanulók körében vizsgálta a papír-ceruza és a számítógépes matematika tesztek itemeinek tulajdonságait. A számítógépes tesztek felvétele a TAO-platformon történt. A nehézségi paraméterek szignifikáns, közepes együttjárást (r = 0,59) mutattak a két adatfelvételi mód között. Elemzése alapján a számítógépes teszt itemei szignifikánsan alacsonyabb diszkriminációs indexszel bírnak, azaz kevésbé jól választják szét a gyengébben és jobban teljesítő tanulókat. Egyéb (tartalmi vagy formai) jellemzők mentén csak az induktív gondolkodás dimenzióban és a grafikus elemet tartalmazó feladatok kapcsán volt kimutatható szignifikáns eltérés. A tanulók neme szerint nem állt fenn számottevő különbség a két médium között. Későbbi, szintén a matematika területére irányuló vizsgálatokban (Hülber & Molnár, 2013; Pásztor-Kovács et al., 2013) egyik vizsgált évfolyamon sem találtak szignifikáns teljesítménykülönbséget a két tesztmédium között.

Herczegné Goldschmidt Zsuzsanna (2016) 4. évfolyamos tanulók körében az eDia rendszerrel végzett elemzése szerint a szövegértés területen pozitív, közepes mértékű együttjárást (r = 0,58) mutattak a papír-ceruza és a számítógépes teszteredmények. Vizsgálata során általánosságban arra jutott, hogy papíralapon a teszt százalékos megoldottsága szignifikánsan, 3–4 százalékkal magasabb, azonban ez a szórásnak kevesebb, mint a negyede, tehát a különbség csekélynek mondható. A tesztek itemszintű vizsgálatakor adódott olyan item is, amely az online teszten bizonyult könnyebbnek.

3.4.4. Az eDia projektek adaptív teszteléssel kapcsolatos vizsgálatai.

A számítógépes adatfelvétel további vizsgálatok előtt nyitotta meg az utat. Egy számítógépes mérés a feladatok kiközvetítésén és a válasz rögzítésén túl képes lehet további adatok rögzítésére és tárolására (Molnár & Csapó, 2019). Ilyen adat lehet a válaszidő, az egérmozgás követése vagy a teszt közbeni navigáció (görgetés vagy a feladat fülek közötti váltás) adminisztrálása. A rendszer ezeket az információkat úgy nevezett log fájlokban vagy napló fájlokban tárolja, és a későbbiekben lehetőség van ezen adatok elemzésére is. A mérés naplófájljainak vizsgálatával lehetőség nyílhat bonyolult feladatmegoldások hátterében álló folyamatok vizsgálatára (pl. Greiff et al., 2015), vagy a válaszidő és a válasz összevetésével a motiválatlan válaszadás detektálására (pl. Csányi & Molnár, 2021), vagy bizonyos item válaszoló általi kihagyásának modellezésére (pl. Lu & Wang, 2020).

Egy másik vizsgálati terület az adaptív tesztelési mód (ld. 2.3 fejezet) bevezetésének megalapozó vizsgálata. Magyar és Molnár (2013) 158 fővel, 5–8. évfolyamos tanulók körében végzett vizsgálatot 2012 őszén, melynek során 28 itemből álló lineáris tesztet hasonlított össze ugyanilyen hosszú, négy szakaszból és három szintből álló (1–3–3–3) többszakaszos adaptív teszttel. Eredményeik alapján az adaptív teszt személyszeparációs reliabilitási mutatója eléri a lineáris teszt megbízhatósági mutatóját, ez alapján alkalmas ugyanannak a jelenségnek a vizsgálatára. A lineáris és az adaptív teszten a kapott eredmények erősen korrelálnak egymással (r = 0,82). A két teszteredmény nem különbözött egymástól szignifikánsan (kivéve a 8. évfolyamot, de a különbség itt is alatta marad a szórás 25%-ának, azaz csekélynek mondható). Az adaptív teszt kezdő szakasza a diákok 40%-át helyesen sorolta be a teljes teszt alapján megállapított szintre. A teszteredmény hibájának nagysága jellemzően alatta maradt a lineáris tesztből számolt hibának, az alacsony és a magas képesség tartományokban a lineáris teszt lényegesen nagyobb hibával mért, mint az adaptív teszt. A három legegyszerűbb tesztútvonal (könnyű-közepes-nehéz) a tesztinformációs görbék alapján több információt szolgáltattak, mint a változatos nehézségű itemeket tartalmazó lineáris teszt. Ez alapján ugyanolyan hosszúságú lineáris és több szakaszos adaptív teszt közül az adaptív minden szempontból legalább ugyanolyan jónak bizonyult, mint a lineáris.

A szóolvasási készség teszt adaptív megvalósításához kapcsolódó vizsgálat (Magyar, 2014a, 2015) 2014 tavaszán 154 tanuló részvételével 1–5. évfolyamokon zajlott. A teszt négy szakaszos öt szintű (1–4–5–5) többszakaszos adaptív szerkezetben valósult meg. Az adaptív teszt megfelelő reliabilitási mutatókkal rendelkezett, valamint jól elkülönítette mind az egyes klasztereket (évfolyamokat), mind az egyes tanulókat a szóolvasási képesség szempontjából. A teszten elért eredmény és a befejező szakasz szintje jól illeszkedett egymáshoz. A harmadik és negyedik szakasz szintje között még nagy számú változás volt (kb. a tanulók 80%-a esetében), ez alátámasztja a szerkezet helyes választását.

A kismintás mérést követően 3 220 tanuló nagymintás vizsgálata a 4. és 5. évfolyamon megerősítette a pilot kutatás eredményeit (Magyar & Molnár, 2015). A számítógépes lineáris teszttel történő összehasonlításban az adaptív teszt reliabilitás-mutatói és a teszt információja magasabb volt, mint a lineáris teszté. A tanulónkénti eredmények erős együttjárást mutattak (r = 0,74, p < 0,01), a pontok és képességszintek között nem volt kimutatható különbség. A mutatók közötti különbségek, a korábbi kutatások eredményeivel összhangban, a képességszintek mentén valamelyest eltértek.

Az eDia fejlesztésének lépései megmutatják a papír-ceruza tesztektől az adaptív tesztelés irányába vezető egyik lehetséges utat, mely valódi adatfelvételekből származó eredmények összehasonlításával értékeli a teszt fejlesztésének egyes elemeit.

3.5. Országos kompetenciamérés

Mivel kutatásom az Országos kompetenciamérés keretrendszerén és adatain alapszik, ezért a következőkben az OKM fő jellemzőit, tartalmi és fogalmi kereteit, valamint fejlődését mutatom be a kezdeti papír-ceruza tesztektől a digitalizáció 2023. évi állapotáig.

3.5.1. Az OKM általános jellemzői

Az Országos kompetenciamérés több céllal indult. Egyrészt a Bevezető fejezetben már említett Monitor mérések hagyományait követve, általános helyzetjelentés biztosítása az oktatásirányítás felé. Másrészt a PISA mérés módszertanához hasonló, az oktatási intézmények felé nyújtott eredményességi mutatókként szolgálnak a kapott számszerű eredmények. Harmadrészt a mérés a nemzetközi mérési kultúra megismertetését és elterjesztését tűzte ki maga elé (Csíkos & Vidákovich, 2012), amit az egyre nagyobb számú nemzetközi mérések is indokoltak. Ezeket a célokat az OKM a 2022. évvel kezdődő digitális országos mérések esetében is megtartotta (Balázsi et al., 2021).

Az OKM 2001-es indulása óta hatalmas fejlődésen ment át, ennél fogva szerepe is folyamatosan változott a hazai oktatáspolitika és neveléskutatás területén. A 2001. novemberi mérés pilotnak tekinthető, mind a mérés időpontja, mind a felmérni kívánt populáció tekintetében. A 2003. évtől kezdődően az OKM időpontja május utolsó szerdája. A mérés alapját képező populációkban is változás történt, 2003-ban a 6. és 10. évfolyam, 2004-től a 8. évfolyam tanulói is részt vesznek a mérésben. 2006 és 2012 között a 4. évfolyamos tanulók körében is történt egy megvalósításában hasonló, de eltérő belső tartalmú mérés Országos készség- és képességmérés elnevezéssel[20]. 2013 és 2021 között ennek a korosztálynak a mérése az iskolák szándéka szerint, önkéntes alapon, a korábbi mérési anyagokkal volt lehetséges[21]. Az évfolyamok mellett bővült a felmért tanulók köre, a kezdeti reprezentatív minta után első lépésben 2006-tól a 8. évfolyamon, 2008-tól a 6. és 10. évfolyamon is teljes körű lett a felmérés, ezzel az iskolák számára történő

[20] https://www.oktatas.hu/kozneveles/meresek/4_evfolyam_keszseg_2012
[21] https://okm.kir.hu/fit/Default.aspx

visszajelzés rendszere teljeskörűvé vált. 2008-tól bevezetésre került a mérési azonosító, mely lehetővé teszi a tanulói adatok nyomon követését.

Az OKM 2008–2021 között változtatás nélkül került megszervezésre, kivéve a 2020. évi mérési kört, amikor a COVID-19 pandémia miatt elmaradt. 2022-ben több nagyobb változtatás történt. Egyrészt az adatfelvétel nem papír-ceruza tesztekkel, hanem számítógépes mérési felületen történt, másrészt kiegészült a természettudományos műveltség mérési területtel (Balogh, Faddiné Buza et al., 2021), harmadrészt egyetlen mérési nap helyett mérési időszakokban került kitöltésre. 2023-ban a 4–11. évfolyamokra bővült a mért évfolyamok köre, bár nem mindegyik évfolyamon és képzési típusban mérik mindegyik területet. A szakképzésben tanulók csak 10. évfolyamon és csak matematika és szövegértés területen vesznek részt a mérésben. 2024-től a mérési területek az 5–11. évfolyamokon kiegészülnek a történelem és digitális kultúra területekkel (Oktatási Hivatal, 2023b).

Az OKM-ben alkalmazott standardizált tanulóiteljesítmény-mérés az iskolák teljesítményének mérésére alkalmas, mivel az iskola szintjén a mért tanulói szintű értékek hibái már kiegyenlítődnek, a mérés eredménye megbízható. Ugyanakkor az iskolai eredmény a kisebb iskolák vagy kislétszámú alcsoportok (pl. halmozottan hátrányos helyzetű tanulók) esetében időben kevésbé stabil, és érzékeny a tanulói összetétel kisebb változásaira. Ez egyrészt az ösztönzők (jutalmak és büntetések) egyenlőtlen érvényesülését okozza, másrészt kedvezhet a mérés manipulálásának (Horn, 2010; Kertesi, 2008; Tóth, 2010).

Az Országos kompetenciamérés indulásakor az egyes mérések eredményeképpen két produktum született: az oktatásirányítás (és az oktatáskutatás résztvevői) felé az országos jelentés, az oktatási intézmények felé az iskolajelentés, mely segítségével tágabb környezetben értékelhetik saját tevékenységüket, objektív adatok alapján összehasonlíthatják magukat hasonló adottságokkal rendelkező intézményekkel (Balázsi et al., 2009). Az iskolák önértékelését[22] és fenntartó általi értékelését segítő iskolajelentések szempontjai az évek során egyre bővültek, a jelentések 2007-től nyilvánosak, bárki számára elérhetők[23]. A mérési azonosító 2008-as bevezetése és a tanulói eredmények korábbi mérési eredménnyel történő összekötése óta (először 2010-

[22] Egy példa, az Ebesi Arany János Magyar-Angol Két Tanítási Nyelvű Általános és Alapfokú Művészeti Iskola önértékelése (http://ebesarany.hu/wp-content/uploads/2020/08/A-2019.-%C3%A9vi-OKM-eredm%C3%A9nyeinek-%C3%A9rt%C3%A9kel%C3%A9se.pdf)
[23] https://www.oktatas.hu/kozneveles/meresek/digitalis_orszagos_meresek/eredmenyek

ben) a mérési azonosító birtokában a tanulók és szüleik elérik a Tanulói jelentéseket[24] is. Bár a mérés eredeti mérési szintje az iskola, még inkább a feladatellátási hely, mivel ezen a szinten az egyéni mérési hibák már kiegyenlítik egymást, a hangsúly egyre inkább az egyéni eredményekre helyeződik (Karkó, 2023), holott jelenleg az egyéni mérési hibák nagysága ezt nem teszi lehetővé (Kertesi, 2008).

A tanulói tesztfüzetek mellett a tanulók és szüleik háttérkérdőívet is kitöltenek (Auxné Bánfi et al., 2014). A Telephelyi, illetve az Intézményi kérdőívek az iskola objektív adottságaira és a vezető szubjektív megfigyeléseire is rákérdeznek. Az Intézményi kérdőívet 2004 óta az oktatási intézmény vezetője tölti ki, és felvilágosítást ad az intézmény pedagógusállományának összetételéről és az OKM eredményeinek hasznosításáról. Végleges formáját 2006-ban nyerte el, ekkor vált külön az intézményi és a telephelyi szint (Auxné Bánfi et al., 2014). A Telephelyi kérdőívet a feladatellátási hely, azaz az egy postai címhez tartozó egység vezetője tölti ki, és információkat szolgáltat az épület állagáról, a tantermek számáról és funkciójáról, a pedagógusállomány összetételéről, az egyes képzési típusok létszámáról és összetételéről. 2019-ben az önálló intézményi kérdőív megszűnt, a kérdések jelentős része átkerült a telephelyi kérdőívbe, és a feladatellátási helyre vonatkozik. (Ez a változás részben a szakképzési centrumok 2015-ben történt kialakítására vezethető vissza, mivel egyetlen intézmény alá a korábbihoz képest lényegesen nagyobb számú tagintézmény (telephely) tartozott (120/2015. (V. 21.) Korm. rendelet, 2015).) A *Tanulói kérdőív* 2006-ban nyerte el közel jelenlegi formáját[25]. A tanulók és szüleik önkéntes alapon adnak információt a tanulmányi előmenetelre, az iskolai életre, az otthoni körülményekre és a szabadidős tevékenységre vonatkozóan. A *Tanulói kérdőív* 2022-től kezdődően *Háttérkérdőív* néven, azonos tartalommal, az OKM digitális felületén kerül felvételre.

A *Tanulói kérdőív* néhány változójából (szülők iskolai végzettsége, otthoni könyvek száma, otthoni számítógép, saját könyv) kialakításra került a családiháttér-index (Auxné Bánfi et al., 2014; Balázsi & Zempléni, 2004), amelybe 2013-ban a halmozottan hátrányos helyzet jelzése is belekerült. 2014-ben az index összetétele és az egyes tényezők súlya felülvizsgálaton esett át. A családiháttér-index számításának bevezetése lehetővé tette a társadalmi különbségek figyelembevételét az iskolai eredmények értékelésében, egyszersmind ráirányította a figyelmet az egyenlőtlen szociokulturális

[24] https://okm.kir.hu/fit2/Jelentes/TanuloKereso

[25] Az egyes mérésekhez tartozó háttérkérdőívek elérhetők az Oktatási Hivatal honlapján: https://www.oktatas.hu/kozneveles/meresek/kompetenciameres/hatterkerdoivek

helyzettel kapcsolatos eredmény-különbségekre. Ez annál is inkább fontos, mivel „A PISA eredményekből az évek során tartósan leszűrhető tanulság, hogy a méltányosság és eredményesség kéz a kézben járnak.” (Lannert, 2015, p.24).

Az eredményesség mérése történhet *keresztmetszeti modell* szerint, amely csak az adott mérés eredményét veszi figyelembe, vagy egy másik lehetséges módja, a hozzáadott pedagógiai érték egyik fajta értelmezés szerinti számítása (Kertesi, 2008). A hozzáadott pedagógia érték modellek családjának egyes elemei arra törekszenek, hogy az iskola nyers eredményéből kiszűrjék a külső tényezők befolyását (Széll, 2018). A modellek mindegyike külső tényezőként veszi számításba a tanuló korábbi teszteredményét, míg a modellek egy jelentős része szűri a tanuló aktuális szociokulturális hátterét is (Nahalka, 2023b). A legegyszerűbb a *növekedési modell* (Balázsi, 2016), amely a korábbi eredményhez képesti változást tekinti eredménynek. Ez a modell jól megfelel az adaptív pedagógiai elméletnek, amely szerint az iskola feladata, hogy minden tanulót a maga sajátosságainak, eltérő családi és kulturális hátterének megfelelően fejlesszen, azaz az aktuális és korábbi eredmény különbségét az iskola tevékenysége következményének tekint (Nahalka, 2023a). A *hozzáadottérték-modell* a korábbi teljesítmény és a kontextuális tényezők alapján várt eredménytől való különbséget, vagyis a reziduálist tekinti eredménynek. A legösszetettebb modellek igyekeznek figyelembe venni a korábbi tanulói eredmények mellett a családi és iskolai ráfordításokat és az egyéni jellemzőket is (Kertesi, 2008). A szociokulturális háttér kiszűrése esetén a magas hozzáadott pedagógiai érték azt jelenti, hogy az iskola tanulói átlagosan magasabb képességpontot értek el, mint a hasonló átlagos tanulói háttérrel rendelkező iskolák tanulói. Ez a megközelítés jobban illeszkedik a deficitelmélethez, mely szerint a tanulási eredmény szoros összefüggésben van a családi-otthoni jellemzőkkel, és az iskola feladata a hátránykiegyenlítés, a hátrányos helyzetből adódó kezdeti lemaradás csökkentése (Nahalka, 2023a).

Az OKM a keresztmetszeti eredmények mellett többféle modell szerint határozza meg az iskolák eredményességét. Az egyik a tanulók átlagos családiháttér-indexéből az átlagos teljesítményre számított regressziós becslés reziduálisa, amely az *iskola hátránykompenzáló szerepét* mutathatja. Mivel ezek a mutatók nem veszik figyelembe a tanulók korábbi eredményeit, ezeket inkább nevezhetjük kontextuális eredményességi mutatónak (Balázsi, 2016). A másik a tanulók két év alatt elért átlagos teljesítményváltozását jeleníti meg, kiszűrve belőle a korábbi tanulói eredményt, tehát hozzáadottérték-modellt alkalmaz. Ez a mutató az *iskola fejlesztő szerepét* jelezheti. Amennyiben magas, az azt mutatja, hogy a telephely tanulói átlagosan magasabb

eredményt értek el, mint a hasonló korábbi átlagos tanulói eredménnyel rendelkező iskolák. Az OKM esetében az átlagostól eltérő fejlődést a tanulók korábbi képességpontjaiból a jelenlegi teljesítményre számított regressziós modell reziduálisa mutatja, tekintet nélkül arra, hogy a tanuló korábban is az iskola tanulója volt-e. Egy harmadik, szintén hozzáadottérték-modell, melyet az OKM-ben *komplex modellnek* neveznek, mind a korábbi eredményt, mind a szocioökonómiai jellemzők némelyikét, mind osztály- vagy telephely szintű iskolai jellemzőket figyelembe veszi az aktuális eredmény előrejelzőjeként, az iskolai eredményességet itt is a reziduálisok jelentik (Balázsi, 2016). Az egyszerű átlagokat tehát az így számolt mutatók alapján hátránykompenzáló, illetve fejlesztő szempontok bevonása egészíti ki, melyekkel azok az iskolák is jó példaként szolgálhatnak, amelyek sikeresen foglalkoznak hátrányosabb helyzetű tanulókkal[26].

Mint látható, az OKM az iskolák értékelése során nagy hangsúlyt fektet a tanulói különbségek feltárására, és az iskolai eredményt a tanulói eredmények összességeként értelmezi. A visszajelző rendszer, elsősorban a Telephelyi jelentések, a keresztmetszeti módszereken túl hozzáadott pedagógiai érték módszerekkel számított indikátorokat is alkalmaz (Horn, 2010; Oktatási Hivatal, 2020), ezért a *Tanulói kérdőív* igyekszik részletesen feltárni a tanulói háttér különbségeit. Ebből következik, hogy az OKM eredményei egyelőre inkább kiszolgálják az iskolai esélyegyenlőtlenségek keletkezésének deficitelméletére alapozott kutatásait, és kevés támponttal szolgának az adaptív pedagógia esélyegyenlőtlenség-elméleten alapuló kutatásaik számára. A Telephelyi kérdőívre a személyi és eszközellátottság, a fegyelem és a motiváció, vagy a továbbtanulási irányok kapcsán kapott válaszok, és azoknak az iskolák eredményességével való összekapcsolása árnyalhatja az iskolák pedagógiai munkájának megítélését. Bár néhány kérdés szerepel a telephelyi és intézményi kérdőívekben az iskolák belső működésére, például a rekrutáció szempontjaira vagy a tanulói csoportok kialakítására vonatkozóan, de a kérdőív nem foglalkozik az iskola pedagógiai programjával, innovációs tevékenységével, szervezeti kultúrájával.

A telephelyi eredményeket az intézményfejlesztésben is felhasználják (Bander et al., 2015). A tanulói képességek esetében megtörtént az alapszint és minimumszint meghatározása, az utóbbi alapján rosszul teljesítő iskoláknak intézkedési tervet kell készítenie (20/2012. (VIII. 31.) EMMI rendelet, 2012). Az iskolák részére a 2021-es

[26] https://www.oktatas.hu/kozneveles/meresek/kompetenciameres/kiemelkedo_teljesitmenyu_iskolak

mérésekig interaktív elemzőprogram[27] áll rendelkezésre, mellyel a pedagógiai munka tervezését segíthetik, például azonosíthatják azokat a feladattípusokat, melyeket iskolájuk tanulói a nekik megfeleltethető tanulócsoportoknál kevésbé sikeresen oldottak meg. Hasonlóan, a középiskolák tanulóik általános iskolai OKM eredményeit elemezve kialakíthatják a tanulócsoportokhoz illeszkedő pedagógiai munka és a lehetséges innovációk terveit. A tanulók eredményeire alapozott elemzések a teljesítménybecslés nagy egyéni hibája miatt akkor tekinthetők érvényesnek, ha telephelyi vagy legalább osztályszintű mutatók kerülnek kialakításra, vagy a mérések a kétévenkénti gyakoriságnál lényegesen sűrűbben történnek[28]. A tanfelügyelet éves munkatervének előkészítését is segíthetik az OKM telephelyi eredményeinek elemzése.

A tanulói, feladatellátási, intézményi és fenntartói adatfájlok – mely, mint közpénzből megvalósuló adatvagyon a kutatók rendelkezésére kell, hogy álljanak – mellett kérésre adatlekérdezések és elemzések készülnek kutatások és fejlesztések előkészítéséhez. A telephelyi eredmények felhasználhatók egyrészt más kutatásokban a mintaválasztás előkészítésében, másrészt az eredmények elemzése és értelmezése során külső indikátorként, esetleg valamilyen beavatkozás eredményességének méréseként[29]. Végső soron a neveléstudomány és a tanulói mérések kölcsönösen támaszkodnak egymásra, ahol „… a mérések a tudomány eredményeire építenek, a vizsgálatok kivitelezése tudományos szakemberek feladata, a konkrét tevékenységet a tudomány által kidolgozott eszközrendszerek használata jellemzi. … a tudomány szerves részeként jelennek meg olyan kutatások, amelyek támaszkodnak a tanulói teljesítmény vizsgálata során született eredményekre.” (Nahalka, 2015, p.23). Így jogos felvetés, hogy az OKM fejlesztése közben a tudomány, akár a neveléstudomány, akár a matematika és mérésmódszertan eredményeire támaszkodjunk. A mérés eredményeinek felhasználhatósága a döntéshozatalban (Szemerszki, 2015), illetve a mérések megítélése (Molnár & Magyar, 2015; Tóth, 2014) önmagukban is neveléstudományi kutatások tárgyát képezik.

[27] https://okm.kir.hu/elemzes/

[28] A Kiskunhalasi SZC Kiskunfélegyházi Kossuth Lajos Szakképző Iskolája és Kollégiuma példája az OKM eredményeinek felhasználására, saját értékelési rendszer és rekrutációs innováció megalapozásához (https://ppk.elte.hu/file/Kiskunfelegyhaza_esettanulm.pdf).

[29] A 2010-es évek végéig a személyiségi jogok megfelelő védelmével szabadon a kutatók rendelkezésére álltak, az utóbbi években az adatvédelmi irányelvek szigorodása miatt a tanulói adatfájlok a mérési azonosító és az OKM-ben álnevesítést lehetővé tevő saját azonosító helyett csak a mérésben részt vevő személyek azonosítását ellehetetlenítő anonim adatfájlok voltak elérhetők. Ezek alkalmatlanok arra, hogy az egyes tanulói sorokat intézményhez vagy más évek adatsoraihoz lehessen kötni, vagyis tanulói adatok alapján sem az intézmények vizsgálata, sem longitudinális elemzések nem voltak lehetségesek. A fenntartói, intézményi és feladatellátási adatfájlok változatlan tartalommal készültek el.

A jelentések mellett 2021-ig minden évben elkészültek a tesztfeladatok elemzési eredményeit tartalmazó *Feladatok és jellemzőik* kötetek is. Ehhez a területhez kapcsolódik az OKM és általában a komplex jelenségeket mérő tesztek egyik kritikája (Nahalka, 2015), miszerint az egydimenziós jelenség mérésére fejlesztett mérőeszköz által nem az eredetileg meghatározott komplex jelenséget (szövegértés, problémamegoldás), hanem annak a mérőeszköz feladatai által jól meghatározott valamilyen egydimenziós változatát mérjük. Nahalka (2015) szerint ez nem probléma abban az esetben, ha a kutató az eredmények felhasználása közben észben tartja a két jelenség – a komplex fogalom és az egydimenziós mérés – közötti különbséget. 2022-től a tesztfeladatok nem nyilvánosak, mivel a papír-ceruza Kiegészítő mérés (Auxné Bánfi et al., 2014) megszűnésével az egyes évfolyamok és egymást követő évek eredményeinek összekötése a több mérési körben szereplő úgynevezett híd feladatok (korábban Core feladatok) segítségével valósul meg (Oktatási Hivatal, 2023a), ennek azonban feltétele, hogy az itemek ne legyenek ismertek.

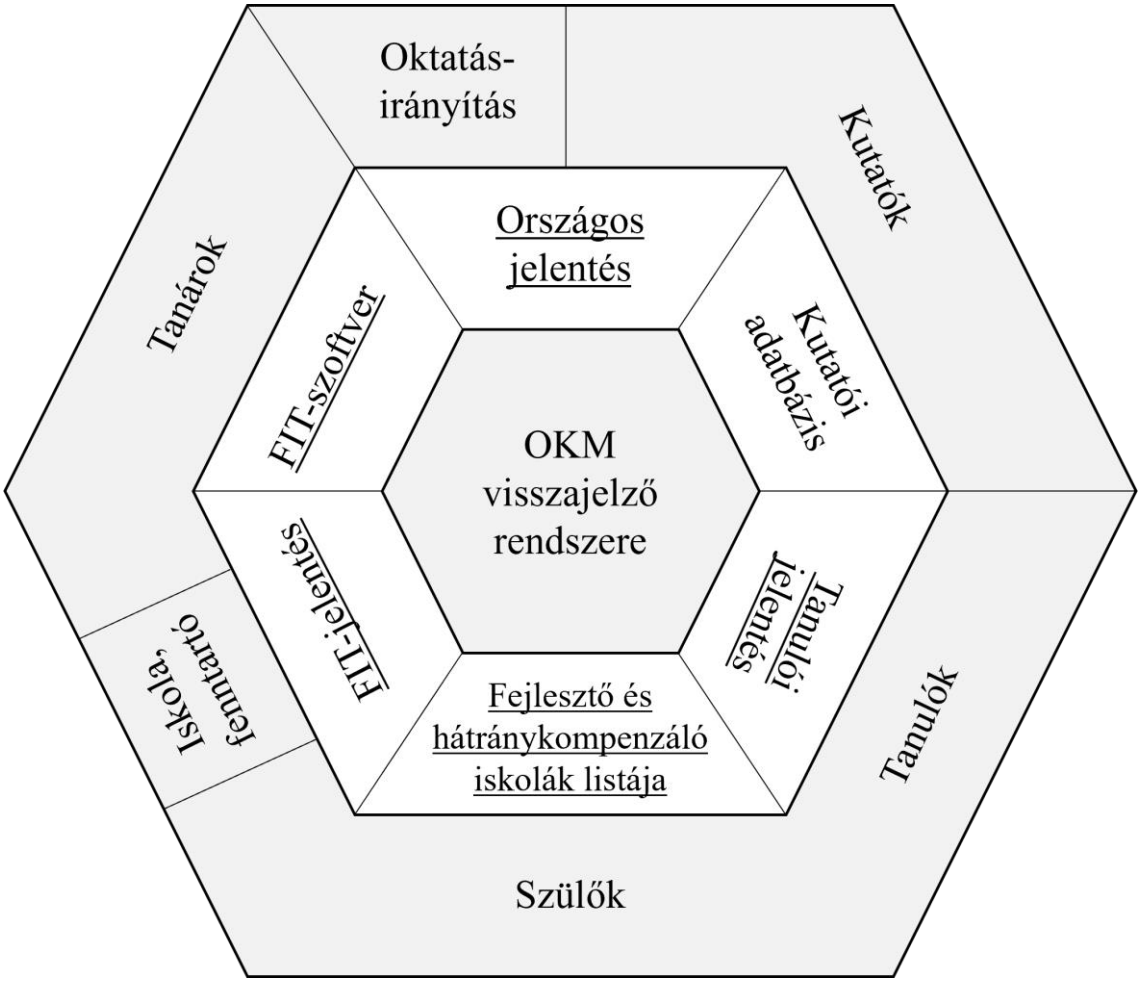
Hasonló kritika, hogyha az iskola komplex működését kívánjuk feltérképezni és értékelni, akkor nem elegendő egyetlen szempontot figyelembe venni. Biesta (2009) szerint egyetlen oktatási cél, a mérhető teljesítmény növelése felé törekvés más területek, például a társadalmi nevelés háttérbe szorulását eredményezi. Az OKM esetében kialakult mérési rendszerről van szó, és bár lehetne kutatásom témája a mérés belső tartalmának, fogalmi háttérnek és a mért jelenségnek a pontos feltárása, tisztázása és fejlesztése, jelen kutatásomban nem erre fókuszálok, hanem a mérési rendszer fejlesztésére. Azzal a feltevéssel fogok élni, hogy az OKM valamilyen egydimenziós jelenséget mér, amely jelenséget a teszt feladatai definiálják, és az egydimenziósságot a tesztfejlesztés biztosítja (Nahalka, 2015). Emellett a későbbiekben kitekintek a komplex jelenségeket mérő többdimenziós modellek alkalmazási lehetőségre is.

A 2021-es méréssel véget ért papír-ceruza OKM eredményei egy visszajelző rendszer keretei között fejtették ki hatásukat (9. ábra), amely rendszer az eredeti célközönségen kívül az iskolák fenntartói felé, a mérési azonosító 2008. évi bevezetése után 2010 óta a tanulók és szüleik felé is szolgáltattak információkat. Ez a komplex tájékoztatás megfelel az elszámoltathatóság elvének, azon belül a nyilvánosság aspektusának (Bander et al., 2015). A nyilvános telephelyi jelentések, a hátránykiegyenlítésben vagy a fejlesztésben kiemelkedő iskolák listái segíthetik az iskolaválasztást. Az egyéni fejlődést tanulói jelentés formájában lehet követni, azonban az egyéni teljesítmény becslésének nagymértékű hibája miatt csak nagy

bizonytalansággal[30]. A telephely fejlesztő szerepének mérését szolgáló regressziós modellek nemcsak a valóságos képességpontokat közlik, hanem tájékoztatnak arról is, hogy a hasonló teljesítményű iskolába/osztályba járó, esetleg hasonló családi háttérrel rendelkező tanulók átlagosan milyen eredményt értek el. Így a tanuló, ismét csak nagy bizonytalansággal értékelheti tágabb közegben a maga és iskolája eredményeit.

9. ábra

Az Országos kompetenciamérés visszajelző rendszere 2021-ig. (Forrás: saját ábra)



Megjegyzés. Az aláhúzott elemek nyilvánosak (pl. Országos jelentés), vagy a megfelelő azonosítók birtokában (pl. Tanulói jelentés) elérhetők.

[30] A mérési hiba azt jelenti, hogy a teszt alapján becsült teljesítmény mekkora bizonytalansággal terhelt. Példa: ha egy tanuló teljesítmény 1570 pont, és a mérési hiba 60 pont, akkor azt állíthatjuk nagy biztonsággal (95%-os valószínűséggel), hogy a tanuló eredménye valahol 1450 és 1690 pont között van, tehát matematikai eszköztudása a negyedik szint alja és az ötödik szint teteje között helyezkedik el.

Az OKM-mel kapcsolatban fontos kritikaként merül fel magának a mérésnek a pontossága. Mivel a mérési azonosítók által lehetséges az anonim és személyre szabott visszajelzés, elvárás lehet, hogy a közpénzből finanszírozott adatfelvétel alanyai, a tanulók is megismerhessék a saját eredményeiket. A tesztek egyéni hibái összemérhetők a tanulói teljesítmény populációs szórásával, mely alapján jól körvonalazható, hogy az eredmények az egyén szintjén nagy bizonytalansággal használhatók. Ez a bizonytalanság a mérés elsődleges fókuszából, az iskolai teljesítmény mérésének céljából származik: a mérés nem a tanulói, sokkal inkább az iskolai eredmény nagyobb fokú pontosságára törekszik, a tesztelméletek szerint pedig az egyéni hibák iskolai szinten kiegyenlítik egymást. Ez az oka, hogy bizonyos számú értékelhető tanulói eredmény nélkül (5 fő) az Oktatási Hivatal nem állítja elő a telephelyi jelentést, mivel annak eredményei iskolai mérési szinten nagy hibát hordoznának. Hasonlóan, bizonyos feltételek szükségesek a regressziós eredmények (hátránykiegyenlítő és fejlesztő szerep) megjelenítéséhez is, ezek egyike, hogy legalább 10 tanuló rendelkezzék a becslés alapjául szolgáló családiháttér-indexszel, illetve két évvel korábbi eredménnyel. Ennek kapcsán új lehetőséget rejt az adaptív számítógépes technológia alkalmazása: egy szabadon felhasználható, rövid, azonnali kiértékelésű teszt nagyobb gyakorisággal kitöltve – ezáltal több mérési pontot biztosítva – összességében biztosíthatja a tanulói eredmények pontosabb mérését vagy trendek számítását (Nahalka, 2023b). További lehetőség a korábbi tesztek eredményeinek, mint kezdő becsléseknek az alkalmazása, ezáltal pontosabb becslés elérése.

A mérések gyakoriságának növelése, a mérés számítógépes felületre történő átültetése és az automatikus értékelésű feladatok arányának növelése nagymértékben javíthatja a feldolgozás sebességét. A feldolgozás időigényes módja egyértelműen nem támogatta a tanulói szintű eredmények felhasználását. Mivel a végső eredmények a mérést követő év februárjában válnak nyilvánossá, az eredmények gyors felhasználhatósága legalábbis kétséges. Iskolák esetében is inkább a trendek, kevésbé a konkrét évfolyam adatai érdekesek, hiszen azok, akik 8. évfolyamosok voltak a teszt megírásakor, az eredmények megjelenésekor már nem az adott általános iskola tanulói (kivéve a 6 és 8 évfolyamos gimnáziumokat). 2022-ig a tesztfüzetek és *Tanulói kérdőívek* visszagyűjtése, kicsomagolása, szkennelése és kódolása körülbelül 3 hónapot, az adattisztítás további 1–2 hónapot vett igénybe. Az adatok elemzésére 2–3 hónapot számolhattunk, végül a jelentéseket publikáló felület és a FIT-szoftver tesztje következett, 1–2 hónap időigénnyel.

Jól előkészített számítógépes méréssel ez a folyamat lerövidülhet. A digitális mérőeszköz a logisztikai feladatok idő és erőforrásigényét csökkenti jelentős mértékben. A számítógépes adatfelvétel és a digitális adminisztráció adatainak összekapcsolása az adattisztítás folyamatában is gyorsulást hozhat. Tovább lépve: a számítógépes mérési módszer az automatikusan kódolható, zárt végű itemeket támogatja, ezért a nyílt végű itemek számának csökkenésével a kódolás időszaka is rövidül. Az adaptív mérési módszer miatt a mérőeszközben bemért, évekig változatlan, ezért meghatározott paraméterekkel (ld. 2.2 fejezet) rendelkező itemek feladatbankja található, a képességpont becslése is megvalósul automatikusan (ez része az adaptív tesztelésnek), ami az elemzési időt is rövidíti. Egy számítógépes mérés esetén jelentős költségtöbblet nélkül kialakítható a jelenlegi kettőnél több tesztfüzetváltozat, adaptív mérési módszer esetén a tanulók még inkább egyéni vizsgálati eszközt kapnak, így a tesztbiztonság is növekedhet.

2022-től az OKM-et számítógépes mérésként szervezik, amely jellemzően számítógép által ellenőrzött kérdésekből állt. Ennek következményeként 2023-tól kezdődően a tanulók a végleges eredmények előtt megismerhetik a feladatok egy része alapján számított előzetes eredményüket[31] (Oktatási Hivatal, 2023a). Ugyanakkor a számítógépes mérésben rejlő előnyöket még nem sikerült teljesen kiaknázni. A feladatok jelentős része, a papír-ceruza mérésekhez hasonlóan, még nem pontosan bemért item, végleges paramétereiket a mérés után kapják meg, ezért a gyorsabb visszajelzés még nem valósult meg. 2023-tól bővült a mért évfolyamok köre, ami a tanulók gyakoribb (évenkénti) mérését jelenti. A mérési területek bővülése szintén további mérési pontokat jelent a tanulók és az iskola szintjén, a szervezőkre nézve azonban megnövekedett terheket.

3.5.2. Az OKM mérési rendszere: tartalmi és fogalmi keretek 2021-ig

A mérés nevében szereplő *kompetencia* fogalma a neveléstudományban gazdag irodalommal rendelkezik (OECD/DeSeCo, 2003; Vass, 2006; ELTE, 2015). Mivel jelen kutatás alapját az Országos kompetenciamérés papír-ceruza mérései adják, kézenfekvő, hogy az OKM matematikai és szövegértési kompetencia meghatározását vegyem alapul. Az OKM tartalmi keretei (Balázsi et al., 2014) leírása alapján a matematikai kompetencián (vagy eszköztudáson) a következőket értjük:

[31] https://www.tehetsegkapu.hu/eredmeny/elozetes

„• az egyénnek azt a képességét, amelynek segítségével megérti és elemzi a matematika szerepét a valós világban;

• a matematikai eszköztár készségszintű használatát;

• az elsajátított matematikai tudás valós élethelyzetekben való alkalmazásának igényét és az erre való képességet;

• a matematikai eszközök használatát a társadalmi kommunikációban és együttműködésben az egyén életkorának megfelelő szinten.” (Balázsi et al., 2014, p.33)

A szövegértés pedig olyan tantárgyközi kompetencia, mely: „…az írott nyelvi szövegek megértésének, használatának és a rájuk való reflektálásnak a képessége annak érdekében, hogy az egyén elérje céljait, fejlessze tudását, képességeit, kikapcsolódjék, sikerrel alkalmazkodjon vagy vegyen részt a mindennapi kommunikációs helyzetekben.” (Balázsi et al., 2014, p.11)

Mindkét területhez hétköznapi helyzetekhez kapcsolódó feladatok tartoznak, amelyeket szövegértés területen szövegtípus (élményszerző, magyarázó, adatközlő) és gondolkodási művelet (információ-visszakeresés, kapcsolatok és összefüggések felismerése, értelmezés), matematika területen tartalmi terület (mennyiségek, számok, műveletek; hozzárendelések, összefüggések; alakzatok, tájékozódás; statisztikai jellemzők, valószínűség) és gondolkodási művelet (tényismeret és egyszerű műveletek; alkalmazás, integráció; komplex megoldások és értékelés) szerint tovább csoportosíthatunk.

A mérés során alkalmazott tesztfüzetek egy szövegértés és egy matematika eszköztudást mérő tesztrészt tartalmaztak, egyenként 2x45 perc időkerettel. A feladat formai szempontból lehet nyílt végű (önálló válaszadás szükséges), feleletválasztós (négy vagy öt lehetséges válasz közül kell kiválasztani az egyetlen helyeset), többszörös választásos (három–öt, jellemzően igaz/hamis kérdés mindegyikére két vagy három lehetőség közül kell a helyeset választani) vagy sorbarendezős (négy vagy öt elem helyes sorrendjének meghatározása). Az utóbbi három típust együttesen zárt végű itemeknek nevezzük.

A feladatok pontozása után a tanuló részére a modern tesztelmélet (ld. 2.2 fejezet) segítségével képességpontot számítanak (Auxné Bánfi et al., 2014). 2009-ig az eltérő évfolyamok eredményeit külön-külön képességskálákon lehetett értelmezni, az egyes évfolyamokat külön trenddel lehetett jellemezni. A képességskálák standardjait úgy

alakították ki, hogy az első mérési évben (a 6. és 10. évfolyam esetében 2003, a 8. évfolyam esetében 2004) a teljesítménypontok populációs eloszlását 500 pont átlagra és 100 pont szórásra transzformálták. A későbbi években (2009-ig) a képességskálák egyezését a korábbi évek képességskáláival az egyes évfolyamokon az biztosította, hogy a kiegészítő mérés itemparamétereit a korábbi évek által meghatározottnak tekintették, az OKM itemparamétereit és a tanulói képességpontokat pedig az ilyen módon rögzített képességskálán határozták meg. A képességbecslést végző programból kinyert 0 várhatóértékű és 1 szórású képességpontokon a bázisév lineáris transzformációját hajtották végre, így a standardizált képességpontok a bázisévben kialakított skálára kerültek (Auxné Bánfi et al., 2014, 6.3 fejezet).

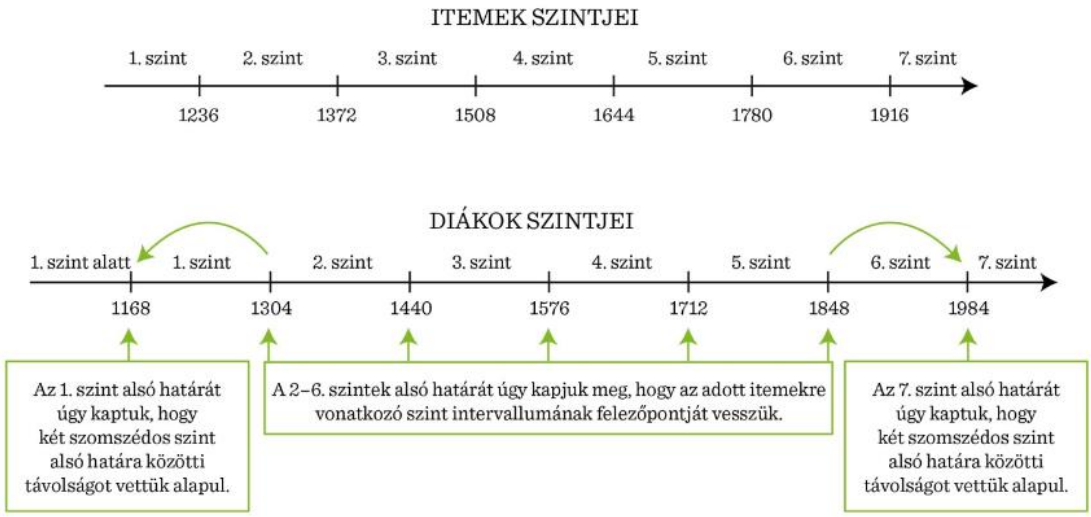
A mérési azonosító és az egyéni visszajelzés 2008. évi bevezetése nyomán 2010-ben volt először lehetséges a tanuló adott és két évvel korábbi képességpontjának összevetése. Ennek feltétele ugyanis, hogy mindkét képességpont azonos képességskálán szerepeljen, ezért 2010-ben az addigi tapasztalatok és a kiegészítő mérés eltérő évfolyamainak tesztfüzeteiben található közös feladatok segítségével közös képességskála került kialakításra. Ezáltal lehetővé vált a különböző évfolyamok eredményeinek összehasonlítása, egyszersmind az egyéni fejlődés mérése. Mivel a 2010. évi eredményeket a 2008. évi eredményekkel kell tanulói szinten összevetni, ezért a közös skála kialakítása a 2008. évig visszamenően megtörtént, ez lett a közös skála bázisév, ahol a 6. évfolyamos teljesítménypontok populációs átlagát 1500 pontra, szórását 200 pontra transzformálták. A többi évfolyam képességpontjaira ugyanezt a transzformációt alkalmazva az évek és az évfolyamok eredményei összehasonlíthatóvá váltak, mérési és egyéni szempontból egyaránt. Az eltérő skálázást, vagyis a korábbi, évfolyamonként 500 pont körüli átlagú skálák helyett az 1500 pont körüli átlaggal rendelkező skála alkalmazását részben az indokolta, hogy egyértelműsítse a módszertani újítást. Ha továbbra is az 500 pont körüli átlaggal rendelkező skála maradt volna, akkor a 2010 előtti és 2010 utáni évfolyam szintű eredmények látszólag könnyen összevethetők maradtak volna annak ellenére, hogy a tartalmuk eltér egymástól.

A képességpontokat más nemzetközi tanulóiteljesítmény-mérésekhez hasonlóan (pl. OECD PISA (OECD, 2019a)) a jobb interpretálhatóság kedvéért a képességskála egyenletes felosztásával hét plusz egy képességszintbe sorolják (Auxné Bánfi et al., 2014) (10. ábra). Első lépésben az itemeket csoportosítják az igényelt gondolkodási művelet nehézsége szerint, így kialakítva hét itemszintet. Ezeknek felelnek meg a képességszintek, melyek a képességfejlettség valamilyen szintjét, bizonyos gondolkodási

műveletek sikeres kivitelezését feltételezik. A legalsó, 1. alatti szint azokat a képességpontokat reprezentálják, melyek a legegyszerűbb műveletek sikeres elvégzését sem valószínűsítik.

10. ábra

A képességskála felosztása itemszintekre és képességszintekre matematika területen (Forrás: Auxné Bánfi et al., 2014, p.107)



3.5.3. Az OKM digitalizációja

A technológiai váltás lehetséges következményeinek feltárása és az átállás előkészítése érdekében készült egy nemzetközi és hazai tapasztalatokat összegző jelentés (Molnár et al., 2015). A médiahatással foglalkozó fejezet rámutat a nemzetközi eredmények sokszínűségére, amit részben a vizsgálatok módszertani háttere, az eszközök, a vizsgált konstruktumok és a minták különbözőségével magyaráz. Elsősorban nemzetközi empirikus kutatások narratív szakirodalmi áttekintése alapján az OKM esetében nem számít jelentős médiahatásra, legfeljebb a jobb teljesítményű tanulók esetében kissé gyengébb eredményre. A jelentés szerint a felső tagozatos és középiskolás korcsoportban az életkor, a nem, a szocioökonómiai státusz és az IKT eszközök használata szerint jellemzően nem mutatható ki jelentős médiahatás.

Az OKM a korábban említett nemzetközi és hazai mérésekhez hasonlóan szintén áttért a számítógépes tesztelésre. A mérés 2022-től teljes egészében számítógépes adatfelvétellel került megszervezésre (Oktatási Hivatal, 2022b). Bővült a feladattípusok száma (kategóriaválasztás, legördülő menüből választás és „fogd és vidd”), valamint a

szabad szöveges választ igénylő típus különvált a szám és szöveg változatokra, ezzel többféle zárt végű feladattípus szerepel (Balázsi et al., 2021). A papír-ceruza Kiegészítő mérés megszűnt, helyette az évek és évfolyamok eredményeinek összekötését lehetővé tevő korábbi Core itemeket digitalizálták, és híd feladatként a mérésbe beépítették (Oktatási Hivatal, 2023a). A Kiegészítő mérés másik funkciója, az új feladatok próbamérése szintén a számítógépes mérés, vagy, főleg az új mérési területek esetében, számítógépes próbamérés során történik (Balázsi et al., 2021; Balogh, Faddiné Buza et al., 2021).

A tesztváltozatok mellett a korábbi *Tanulói kérdőívvel* tartalmilag megegyező *Háttérkérdőív* kitöltése is számítógépen történt. A 2022-es mérési körben a mérésben résztvevő tanulók mindössze 38–44 százalékára számítható a családiháttér-index (Oktatási Hivatal, 2023a), míg ez a szám 2019-ben 80% (Oktatási Hivatal, 2020), 2021-ben 81% volt (Oktatási Hivatal, 2022a). A korábban önálló idegennyelvi és célnyelvi mérés 2022-től ugyanebben a mérési rendszerben került megszervezésre (Balázsi et al., 2021; Balogh, Garay-Madarász et al., 2021; Oktatási Hivatal, 2022b). 2023-ban a mérésben résztvevők köre bővült, ezen időponttól kezdve a 4–11. évfolyamokon zajlanak a mérések. A mérési területek és a felmért évfolyamok számának jelentős növekedése jól jelzi a papír-ceruza tesztek szervezéséhez képesti kapacitás és a számítógépes tesztek iránti megrendelői igény növekedését.

Mivel az OKM a nemzetközi tanulóiteljesítmény-mérésekhez hasonlóan közöl trendeket, illetve összehasonlításokat, ezért szükséges a médiahatás nagyságának és az eredmények értelmezésére gyakorolt esetleges hatásának vizsgálata. A mérés szervezői támaszkodnak a nemzetközi mérések és a mérések szervezése során szerzett tapasztalatokra (Oktatási Hivatal, 2023a), ugyanakkor a kézirat lezárásáig a médiahatást vizsgáló elemzést az Oktatási Hivataltól nem találtam.

4. Kutatási célok, kérdések

Kutatásomban az Országos kompetenciamérés fejlesztésének a digitalizálásban rejlő egyik irányát, a számítógépes adaptív teszt módszertani kérdéseinek vizsgálatát tűztem ki célul. Olyan megelőző vizsgálatok elvégzését, melyek előkészítik az egyik hazai tanulóiteljesítmény-mérési rendszer esetében a szakmai és mérésmódszertani szempontból sikeres papír-ceruza – számítógépes – adaptív számítógépes adatfelvétel átmenetet.

Kutatásomban megvizsgálom a papír-ceruza és egy lehetséges számítógépes adatfelvétel közötti belső egyezést (a tesztfeladatok oldaláról), és a mérőeszközök eredményeinek megfeleltethetőségét a képességpontok tekintetében. Az automatizált adaptív mérés esetében a teszt következő kérdésének kiválasztási feltétele, hogy a korábbi válaszok alapján azonnal képességpontot számítsunk, ezért csak automatizáltan értékelhető feladatok kerülhetnek a tesztbe. A hosszabb szöveges választ igénylő nyílt végű feladatok elvesztésének a mért konstruktumra és a képességpontok számítására vonatkozó következményei validitási elemzésekkel vizsgálhatók (Arensman et al., 2016; Graham et al., 2019). Következő lépés a számítógépes környezetben megvalósuló adaptív eljárások vizsgálata, a lehetséges módszerek közül az OKM matematikai eszköztudás terület mérési céljaihoz és eszközeihez leginkább illeszkedő lehetőségek kiválasztása. Mint a következő, *A kutatás módszertana* című fejezetben szerepel, első feladat a mérésnek megfelelő mérési struktúra (CAT vagy MST) kiválasztása, majd CAT esetében a soron következő item kiválasztási szabályainak és a teszt megállítási kritériumainak meghatározása. A szövegértés terület, illetve az ahhoz jobban illeszkedő MST esetében a tesztszerkezet (szakaszok és szintek) és a mérés elemeinek (blokkok) kialakítása túlmutat jelen dolgozat keretein.

Kutatásomban az alábbi fő- és alkérdésekre keresem a választ.

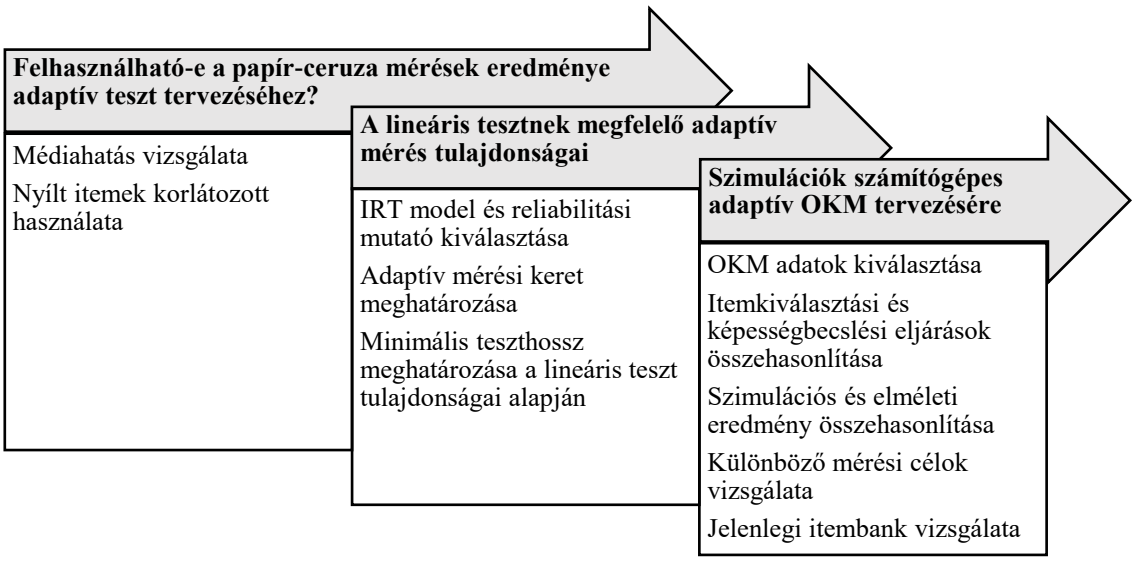
1) Az OKM papír-ceruza méréseiből származó adatok relevánsan felhasználhatók-e a számítógépes adaptív mérés tervezésére?
   a. Mi a számítógépes mérési környezet hatása a mérés eredményére? Kell-e médiahatásra számítani, és ha igen, hogyan kezelhető? (Fishbein et al., 2018) (6.1 fejezet)
   b. A nyílt végű itemek elhagyása mellett is azonos marad-e az OKM mérés tartalmi kerete? Kizárólag zárt végű itemeket alkalmazva milyen eltéréseket tapasztalnánk a tanulók képességpontjának becslésében? (6.2 fejezet)

2) Az eredeti méréssel megegyező mérési pontosság mellett mi az adaptív teszteléssel elérhető legrövidebb teszthossz? (Weiss, 2011) (6.3 fejezet)
3) Az OKM papír-ceruza méréseiből származó adatok alapján mely adaptív mérési elemek valószínűsítik a mérés céljának sikeresebb megvalósítását (a matematikai eszköztudás területen)? (Thompson & Weiss, 2011) (6.4 fejezet)
   a. A papír-ceruza teszttel azonos itemszám mellett az adaptív teszt esetében csökken-e a tanulói képességfejlettség-becslés hibája? (6.4.1 fejezet)
   b. Az OKM adatain alapuló, számítógépes adaptív tesztet imitáló szimulációk alátámasztják-e, hogy lényegesen rövidebb idő alatt (kevesebb itemmel) a papír-ceruza teszt pontosságának megfelelő pontossággal meghatározható a diákok képességpontja/teljesítménye? (6.4.2 fejezet)
   c. Milyen megállítási kritériumok milyen mérési céloknak felelnek meg az adaptív OKM kapcsán?
   d. A megállítási kritériumok között van-e hierarchia, azaz léteznek-e olyan erős kritériumok, melyek teljesülése magával hozza a gyengébb feltételek teljesülését?
   e. Az első 5–10–15–20 kérdés után változik-e még a diák teljesítménye? 5–10 kérdéses teszthossz mellett milyen teljesítménybecslések várhatók?

A kutatás főbb kérdéseinek és feladatainak kapcsolódását a 11. ábra mutatja be.

11. ábra

A fő kutatási kérdések és a hozzájuk tartozó feladatok



5. A kutatás módszertana

Kutatásom egy már létező teljesítménymérési rendszer, az Országos kompetenciamérés következő fejlődési lépcsőjének megalapozó vizsgálata, ennek értelmében alkalmazott kutatás, módszertanát tekintve elsősorban kvantitatív módszertanú. A kutatás részét képezi a számítógépes adaptív mérési technológia elméleti vizsgálata a tanulói képességfejlettség becslésének mérési hibájával kapcsolatban, a médiahatás vizsgálatát pedig szisztematikus szakirodalmi áttekintéssel végeztem, mely kvalitatív módszertanú vizsgálat. A nyílt tételek elhagyása és a különböző item kiválasztási és képesség becslési eljárások összehasonlítása kvantitatív és empirikus jellegű vizsgálatok.

Az 1) kutatási kérdés (és két alkérdése) a matematika területnél bővebben, a szövegértés és természettudomány területeket is bevonva került vizsgálatra. Mivel a papír-ceruza teszt szerkezete, az itemek típusai és ezek aránya hasonló mindhárom területen, valamint az egyes területeken elért képességpont jellemzően nagyon magas együttjárást mutat, ezért a matematikától eltérő területek tapasztalatai nagyobb érvényességet biztosítanak a matematika terület eredményeinek értelmezésekor, egyszersmind rávilágíthatnak a területek közötti különbségekre. A 2) kutatási kérdés általános, mérési területtől független kérdés, tehát nincs szükség sem a matematika területre történő szűkítésre, sem a vizsgált területek számának bővítésére. Éppen ezért a példákat általánosan, a mérés volumene és elvárt jellemzői alapján fogalmaztam meg. A 3) kutatási kérdés kifejezetten a matematika területre vonatkozik, ezért az itt felsorolt eredmények is erre a területre érvényesek.

Az adaptív mérések vizsgálatai jellemzően három módszertani kategóriába sorolhatók. Az *első típus* a matematikai jellegű elméleti levezetések csoportja. Ezek során új matematikai modellek levezetése és elméleti jellegű vizsgálatok történnek. Ilyen lehet például a válaszidőnek, mint dimenziónak a bevonása a válaszmintázat kialakításába (Finkelman et al., 2014; Lu & Wang, 2020).

A *második típust* az empirikus vizsgálatok testesítik meg, amelyek kvantitatív módszertannal valós adaptív tesztek eredményeit vizsgálják és vetik össze lineáris teszteken kapott eredményekkel. Ideális esetben a méréseket kitöltők megegyeznek. Ilyen vizsgálatra példa a szóolvasás képesség teszt adaptációja az eredeti keretrendszer szerint (Magyar, 2014a). Szintén valódi adatfelvételen alapuló vizsgálat tipikus példája a PROMIS (Patient-Reported Outcomes Measurement Information System), egy adaptív egészségügyi mérőrendszer itembankjának fejlesztése és validálása (pl. Crins et al.,

2018). A két módszer közötti átmenetet, *harmadik típusként*, a szimulációs vizsgálatok jelentik.

5.1. Szimulációs technikák

Adaptív mérések esetében a *szimulációs módszerek* olyan technikák, amelyek a nagy számítógépes kapacitáson és az IRT modellek formális matematikai egyenletein alapulnak. Lényege, hogy az előre meghatározott tanulói elméleti képességfejlettség és az itemparaméterek segítségével az alkalmazott IRT modell alapján a számítógép a helyes válasz valószínűségét kiszámítja, és egy 0 és 1 közötti véletlen számmal összehasonlítva szimulálja a helyes vagy helytelen választ. A képességpont becslése a szimulált válaszok alapján történik. A szimuláció előnye a valódi adatfelvétellel szemben, hogy nagyobb mintaelemszámra és számos különböző kondíció összehasonlítására ad lehetőséget (pl. Şahin & Weiss, 2015). A szimulációk szintén tovább csoportosíthatók aszerint, hogy mekkora mértékben használnak valós adatokat.

A *Monte Carlo szimulációk* (Kehl, 2012) teljes egészében véletlenszám generátorral készült adatokat használnak. Ezáltal olyan vizsgálati feltételek is ellenőrizhetők, melyeket valódi adatfelvételben nem lehet garantálni, ugyanakkor az ilyen vizsgálatok érvényessége nagyban függ attól, hogy a generált körülmények mennyire jól követik a valóságot, illetve milyen minőségű a véletlenszám generátor (Araci & Tan, 2022). Sari (2020) a Monte Carlo eljárások közé sorolja azokat a szimulációkat is, ahol az itemek paraméterei valamilyen létező, jellemzően lineáris teszt itembankjából származnak, de a képességfejlettség és a válaszok generáltak. A képességpontok ekkor valamilyen véletlen eloszlásból – jellemzően normális eloszlásból – származó generált adatok (Yang & Reckase, 2020). Lehetséges olyan elrendezés is, ahol a képességfejlettségek egy papír-ceruza vagy nem adaptív számítógépes mérésből származnak, de az itemparaméterek generáltak (Jewsbury & van Rijn, 2020). Ezek az eljárások jól alkalmazhatók olyan esetekben amikor különböző tulajdonságú itembankokat, módszereket vagy szituációkat szeretnénk összehasonlítani, esetleg elméleti levezetések, modellek hatékonyságát szeretnénk bemutatni, a képességfejlettségek eloszlásának pontosabb modellezésével.

Post-hoc szimulációról akkor beszélhetünk, amikor minden item és minden kitöltő esetében rendelkezésre áll az itemre adott valódi válasz (Sari, 2020). Ilyen szimulációk során az elméleti képességfejlettséget a korábbi tesztből származó

képességbecslés reprezentálja, az itemparaméterek szintén a papír-ceruza vagy nem adaptív számítógépes, vagyis lineáris mérésből származnak, valamint a kiválasztott itemre adott válasz sem generált, hanem az adott kitöltő valódi válasza. A szimuláció véletlen eleme a következő item kiválasztása, illetve eltérő képességbecslési módszerek különböző közbülső eredményt és tesztutat eredményezhetnek. Ezek a szimulációk állnak legközelebb az adaptív tesztelés valódi adatfelvétellel történő empirikus vizsgálatához.

Hibrid szimulációk esetén részben átfedő tesztek eredményei állnak rendelkezésre, vagyis egy-egy kitöltő az itembank egy előre meghatározott részével találkozik, jellemzően lineáris teszt formájában. Szintén ez a helyzet akkor, ha a válaszok egy része valamilyen okból elveszett, és pótolni kell azokat. Ilyen vizsgálatokban a válaszok egy része valódi, más része pedig az IRT modell alapján generált. Ugyanakkor az itembank és a tesztből számított képességpont, a post-hoc szimulációhoz hasonlóan, rendelkezésre áll. Hibrid és post-hoc szimuláció esetében jellemző kérdés, hogy azonos itembankon alapuló adaptív teszt ugyanezekkel a kitöltőkkel mennyire jól közelíti a lineáris teszt eredményét.

A szimulációs módszerek eredménye matematikai értelemben nem bizonyítás, ahogy az empirikus eredményeket sem tekintjük megdönthetetlennek. Ugyanakkor az adaptív tesztek tervezése és készítése, az itembank összeállítása költséges és időigényes folyamat. A szimulációs vizsgálatok a tervezés szakaszában kiválóan használhatók az adaptív teszt szerkezeti elemeinek vizsgálatára, legyen szó akár CAT, akár MST szerkezetű tesztről. A Monte Carlo szimulációkat jellemzően a tervezés kezdeti fázisában, míg a hibrid és post-hoc vizsgálatokat a folyamat végén, az elemek végső kialakításában lehet jól alkalmazni (Araci & Tan, 2022; Thompson & Weiss, 2011). Egy a korábbi lineáris teszt vagy tesztsorozat adaptálása során lehetőség van a lineáris tesztek itemeiből álló itembank vizsgálatára, a későbbi item bank hiányzó elemeinek azonosítására. Az egyes elemek (képességbecslés, itemkiválasztás, megállítási kritérium stb.) megváltoztatásával összehasonlíthatjuk a különböző tesztelemek eredményességét.

5.2. Adatforrás és az elemzéshez használt adatok

A kutatáshoz az Országos kompetenciamérés tesztfüzeteinek tanulói és itemszintű adatait használtam fel. A mérés részletes módszertani bemutatása szerepel az OKM Technikai leírásában (Auxné Bánfi et al., 2014). A 2021. évi mérésig a tesztfüzetek, javítókulcsok, illetve a pontozást és az itemek részletes elemzését tartalmazó *Feladatok és jellemzőik*

kötetek mérési évenként és évfolyamonként megtalálhatók az Oktatási Hivatal honlapján[32]. A tesztfüzetek kérdéseire adott válaszok kódjai a kutatói jelentésfájlokban szerepelnek, de a kódok helyett az elemzésbe a válaszok pontszámát vontam be.

Az OKM elérhető adattáblái közül a 2008–2019 évek méréseit elemeztem, a következő okokból:

1) A mérés 2008-tól kezdődően teljes körű a feldolgozás szempontjából, azaz nagyszámú tanulói eredmény érhető el az elemzésbe bevont évekre.
2) A közös képességskála 2010-es bevezetése óta (Auxné Bánfi et al., 2014) az egyes évek mérései mellett a különböző évfolyamok eredményei is összevethetők. A skálát 2008-ig visszamenően alakították ki, a standardizálás bázisa a 2008. évi 6. évfolyamos populáció eredménye (Auxné Bánfi et al., 2014; Oktatási Hivatal, é. n.). A közös skálás eredmény a tanulói fejlődés követéséhez (vö. mérési azonosító) az eredmények közlésében csak 2010-től jelenik meg az Országos jelentésben és a *Feladatok és jellemzőik* kötetekben. A közös skála azóta változatlan, beleértve a 2022. évi számítógépes mérési ciklust is. A PARSCALE programból (DuToit, 2003; Muraki & Bock, 1991) származó, transzformáció előtti képességpontok[33] használatával több év adatait is fel lehetett volna használni, de az évfolyamonként eltérő képességskálák megnehezítik a 2004–2007. évekből származó eredmények összehasonlítását a közös skálát használó évek eredményeivel. A standardizált képességpontok használata azért is előnyös, mert így a szimulációs vizsgálatok eredményét szintén az OKM közös skáláján értelmezhetjük.
3) A közös skála bevezetése az itemparaméterek egységes skáláját is jelenti a modern tesztelméletnek megfelelően, ami szintén előnyös a szimulációs eljárások szempontjából.

Az OKM tanulói populációjáról a nemzeti köznevelésről szóló 2011. évi CXC. törvény 80. § (1) bekezdése és a szakképzésről szóló 2019. évi LXXX. törvény 19. § (5) bekezdése rendelkezik, a mérésben az adott évben 6., 8. vagy 10. évfolyamos tanulók vesznek részt. A mérés tehát teljes körű, bizonyos mentesítő körülmények mellett, amik

[32] https://www.oktatas.hu/kozneveles/meresek/kompetenciameres/feladatsorok

[33] „…a már rögzítettnek tekinthető itemparaméterek felhasználásával a program becslést ad a tanulók képességparaméterére, ami közel 0 átlagú és 1 szórású értékeket jelent. A képességparaméterek standardizálásával (lényegében egy lineáris transzformációval) nyerjük a FIT-jelentésekből ismert 1500 körüli tanulói képességpontokat.” (Auxné Bánfi et al., 2014, 98) A transzformáció utáni változatot standardizált képességpontnak nevezzük.

alapvetően háromfélék lehetnek: (1) sajátos nevelési igényű (SNI) tanulók közül a mozgásszervi vagy érzékszervi (látási, hallási) fogyatékos, a beszédfogyatékos, az autista és az értelmi fogyatékos tanulók; (2) ideiglenes sérülés (pl. kéztörés) miatt mentesíthető tanulók[34]; és (3) nyelvi szempontból mentesíthető (pl. nem magyar anyanyelvű) tanulók.

A tanulói eredmények bevonása tervezett kutatásomba azon alapul, hogy a tanuló az országos eredmény számításába beleszámított-e. Ilyen módon kizárásra kerülnek:

- az SNI miatt mentesülő tanulók (1-es kód),
- a nyelvi okokból mentesülő tanulók (3-as kód),
- a mérést nem magyar nyelven kitöltők,
- egyéb pszichés fejlődési zavarral küzdő tanulók.

Azok a tanulói sorok is kizárásra kerültek az OKM adafájljaiból, amelyekhez üres tesztfüzet tartozik, mivel ebben az esetben nincsenek tanulói válaszok és képességpontok, amiket a szimulációs elemzésbe bevonhatnánk. Fentieken kívül egy külön kizárási feltételt is alkalmaztam: kizárásra kerülnek mindazok a tanulói sorok, ahol a Jelenléti ív alapján valamelyik tesztrészen a tanuló késett, hiányzott, vagy a teszt kitöltését félbehagyta. Erre azért van szükség, mert bár a kitöltött itemek elegendőnek bizonyulhattak az egyik tesztrész elhagyásával, azonban a szimulációhoz kevés a megválaszolt itemek száma, a szimuláció nem fejeződik be vagy fals eredményre vezet. Mivel ezen tanulók eredményei nem kerülnek felhasználásra az elemzésekben, az adattakarékosság elve szerint az adatsoraikat sem mentettem ki.

Összességében, bár a jelentésre jogosult tanulók körének meghatározása az évek során apróbb változásokon ment keresztül, az OKM módszertana 2008 és 2019 között már nem változott jelentősen, így az elemzésbe bevont adatok az adatfelvétel, feldolgozás és elemzés szempontjából homogénnek tekinthetők.

A szimulációs elemzésekbe a 2008–2019. évek OKM méréseinek matematika itemeit vontam be. Ez egyben azt is jelenti, hogy a tesztfüzetekből csak a matematika tesztrészt használtam fel az elemzéshez. Ennek az az oka, hogy a számítógépes adaptív tesztek módszerei közül a CAT technikáját vizsgáltam, azonban a szövegértés teszt szerkezetéhez az MST technika jobban illeszkedik. A matematika feladatok egy vagy két-három egymástól függetlenül megoldható itemből állnak, amihez jól illik a CAT eljárása. Az itemek közül kizártam azokat, amelyek az OKM itemparamétereket meghatározó

[34] 2022-től a mérési időszak bevezetése miatt az ideiglenes testi sérülés miatti mentesség megszűnt, ezek a tanulók pótmérésen vesznek részt.

elemzési fázisa során nem illeszkedtek a méréshez, itemparamétereik nem kerültek meghatározásra, ezért a képességpontok számításához sem használták fel őket.

A tanulói válaszok pontszámán kívül semmilyen más jellemzőt nem használtam fel a kutatásban. Bár egy lehetséges kutatási irány lehetett volna az adaptív teszt viselkedésének vizsgálata különböző tanulói csoportokban (pl. nemek, régiók, iskolatípusok vagy halmozottan hátrányos helyzet szerint), azonban ez a kérdés már túlmutat jelen kutatás keretein. Ezen kérdés vizsgálatához az adaptív rendszeren valódi adatfelvétellel történő empirikus vizsgálati módszer jobban illeszkedne.

5.3. A szimulációs elemzésekhez használt programcsomag (catR)[35]

Az adatok leválogatása, tisztítása és kimentése IBM SPSS programban történt. A CAT eljárás szimulációját (a–e kutatási kérdések) az R (R Core Team, 2016) környezetben működő, nyílt forráskódú programcsomag, a catR (Magis et al., 2017b) segítségével végeztem, ami ingyenes felhasználású és programozható, azaz adaptálható a kutatáshoz, a számítógépes adaptív teszt szimulációjának elvégzéséhez. A tanulói képességpontok és a hozzájuk tartozó hiba becslésére két- és háromparaméteres IRT modellt használtam, mivel az OKM ezt az utóbbit modellt alkalmazza. A szimulációkat elsősorban Monte Carlo, másodsorban hibrid megvalósítással végeztem el. A szimulációba generált és a tanulók tesztfüzeteiből az itemekre adott valódi válaszokat vontam be. A szimulációk futtatáshoz az R (R Core Team, 2016) 4.2.2 és az RStudio (RStudio Team, 2020) 2022.07.2+576 verzióját használtam.

A szimuláció futtatására alkalmas catR programcsomagot először 2011-ben publikálták (Magis és Raîche, 2011). A következő évben megjelent a programcsomag elméleti hátterének és funkcióinak részletes leírása (Magis és Raîche, 2012). 2017-ben a programcsomag egy nagyobb frissítésen esett át (Magis et al., 2017c). Ettől fogva a dichotóm itemek mellett a több pontos itemeket is tudta kezelni, új itemkiválasztási módszerek lettek elérhetők és egy lépésben több tesztalannyal teljes teszt szimulációjára lett képes.

Ugyanebben az évben megjelent az a kézikönyv (Magis et al., 2017b), amely bemutatta a catR és a többszakaszos adaptív tesztelést szimuláló mstR programcsomag funkcióit. A könyv 1) összeveti a lineáris és adaptív tesztek főbb jellemzőit, 2) bemutatja a modern tesztelméleti modellekkel kapcsolatos tudnivalókat, majd 3) ismerteti a

[35] A catR programcsomag bemutatása a (T. Kárász, é. n.) kézirat alapján készült.

számítógépes adaptív és a többszakaszos adaptív mérés elméleti hátterét, 4) példákkal illusztrálva bemutatja a programcsomagot alkotó funkciókat, függvényeket, és 5) részletes, kódokkal, outputokkal és grafikonokkal illusztrált példagyűjteményt ad. A kézikönyvhöz megjelent recenziók (Choe & Fu, 2018; Rutkowski & Valdivia, 2020), melyek kiemelik a könyv logikus felépítését, hiánypótló szerepét az elmélet és a gyakorlat közötti kapcsolat megteremtésében, valamint kiváló példáit és részletes leírásait, ami a témában kezdő olvasóknak széles elméleti áttekintést és jó gyakorlati alapokat biztosít.

A catR programcsomagban megadhatunk vagy generálhatunk itembankhoz paramétereket, a dichotóm itemekhez 1–4 paraméteres modellek szerint, a többpontos itemekhez hatféle különböző modellel. Kiszámíthatjuk adott képességpontban az item megoldási valószínűségét és az item információját, valamint generálhatunk adott item bankhoz és képességfejlettséghez válaszmintázatot. Adott válaszmintázat és itemparaméterek esetén számítható képességbecslés és annak hibája. Kétféle szimulációt alkalmazhatunk: az egyik változat egyetlen képességpont esetén különböző beállítások mellett egy tesztet, a másik több tesztalany esetén teljes mérést szimulál, ahol már az itemek kitettségének szabályozása is lehetséges. Az itemek kitettsége azt jelenti, hogy egy adott itemet a tesztben résztvevők mekkora része ismerhet meg, és a tesztbiztonság szempontból van jelentősége. A programcsomagban hatféle képesség becslési módszer és tizennégy itemkiválasztási módszer érhető el. Négyféle megállítási kritérium alkalmazható, melyek kombinálhatóak is egymással. A szimulációs vizsgálatok során (ld. 6.4 fejezet) háromféle képességbecslési módszert (maximum likelihood, Bayes és expected a-posteriori), három itemkiválasztási eljárást (maximum Fisher-információ, legközelebbi nehézség, legközelebbi maximális információ) és két megállítási kritériumot (meghatározott számú item, meghatározott nagyságú képességbecslési hiba) hasonlítottam össze.

6. Eredmények

6.1. Papír-ceruzáról számítógépes adatfelvételre: médiahatás vizsgálat[36]

Első kutatási kérdésem arra irányult, hogy a papír-ceruza tesztek eredményei mennyiben alkalmasak egy számítógépes adaptív teszt vizsgálatára, azon belül egy feltételezett számítógépes mérés szimulációjára, vagyis a) azonos-e a két vizsgálati konstruktum és b) az itemek jellemzői mennyiben különböznek a két médium esetében.

Az OKM esetében nincs ilyen jellegű vizsgálatról elérhető publikáció, ezért ezt a kérdést a PISA, PIRLS és TIMSS mérések digitalizációjával, azon belül a médiahatással kapcsolatos hazai és nemzetközi tudományos publikációk és mérési dokumentumok szisztematikus szakirodalmi áttekintésével (Rother, 2007) vizsgáltam. A kutatás során a szisztematikus áttekintések és metaanalízisek esetében ajánlott irányelveket (Preferred Reporting Items for Systematic reviews and Meta-Analyses, PRISMA, Page et al., 2021) követtem, azaz az adatbázisokban folytatott keresés célja a lehető legtöbb és legrelevánsabb forrás felfedése és szintetizálása előre jól meghatározott keresési és kizárási kritériumok alapján. A keresést 2021. december 2-án hajtottam végre.

6.1.1. Adatbázisok

A hazai szerzők által publikált eredmények keresésére három adatbázist használtam. Az egyik a magyar folyóiratok tartalomjegyzékeinek kereshető adatbázisa (MATARKA), a másik a Magyar Tudományos Művek Tára (MTMT), a harmadik pedig az Arcanum Digitális Tudománytára[37] volt. A nemzetközi adatbázisok esetében az oktatási, neveléstudományi témájú forrásokat tartalmazókat vizsgáltam. A választást az intézményi hozzáférés befolyásolta. A keresésbe bevont adatbázisok az EBSCO (kivéve a Green és Legal adatbázist), az ERIC, a JSTOR, a ProQuest, a Science Direct és a Web of Science voltak. Ezek mindegyike lehetővé teszi a részletes keresést, összetett logikai kifejezések szerinti keresést, a megjelenés évére és a lektorált (*peer review*) forrásokra történő szűkítést.

A nemzetközi mérések dokumentumait a szervezők oldaláról gyűjtöttem össze. A TIMSS és PIRLS mérések szervezője megegyezik, a dokumentumok az alábbi felületről

[36] A fejezet az *Iskolakultúra* folyóiratban megjelent cikk (T. Kárász & Széll, 2023) alapján készült.
[37] arcanum.com\hu

érhetők el: https://timssandpirls.bc.edu/isc/publications.html. A PISA méréshez kapcsolódó publikációk két helyen, a szervező saját oldalán (https://www.oecd.org/pisa/publications/) és a szervező publikációit tartalmazó felületen (https://www.oecd-ilibrary.org/education/pisa_19963777) is elérhetők. A vizsgálat során az előbbit használtam, mivel itt mérésenként összegyűjtve szerepelnek a kapcsolódó információk. A magyar nyelvű mérési dokumentumokat az Oktatási Hivatal honlapjáról[38], a mérések saját aloldaláról gyűjtöttem le, de nem kerültek a vizsgálatba azok az egyéb elemzések, amelyek más aloldalakon nem lektorált tudományos közlemények formájában jelentek meg. Elfogadtam más országok mérési dokumentumait is, ha azok angol vagy magyar nyelven jelentek meg, és szerepelnek valamely nemzetközi adatbázisban.

A dokumentumok jellemzően a mérések keretrendszerét, a mérés eredményeit és a technikai részleteket közlik. A keretrendszerek azonosítják az egyes méréseket, azok céljait, a méréshez kialakított elméleti hátteret és előre vetítik a mérés bizonyos módszertani és technikai jellemzőit. Az eredmények jellemzően a mérés utáni évben jelennek meg, és az egyes országok eredményei, valamint a nemzetközi eredmények a háttérkérdőívek alapján számított statisztikák mellett tartalmaznak bizonyos mérésmódszertani leírásokat is. A technikai leírások kifejezetten a mérés és elemzés megismerésének és megismételhetőségének érdekében készülnek, és tartalmaznak minden olyan információt, ami az itemek fejlesztésétől az adatok rögzítésén és tisztításán át az elemzés elméleti hátteréig releváns lehet.

6.1.2. Beválogatási és kizárási kritériumok

A keresési időszak kezdetét a számítógépes adatfelvételek időpontjához igazítottam. Mivel a mérés médiumának cseréje már mind a három nemzetközi mérés esetében megtörtént – a PISA 2015-ben, az eTIMSS 2019-ben, a digitalPIRLS 2021-ben alkalmazta teljeskörűen ezt a megvalósítást –, ezért feltételeztem, hogy mind leírások, mind kapcsolódó független vizsgálatok eredményei is rendelkezésre állnak már a témában. Azt is feltételeztem, hogy a 2015-ös PISA mérés előkészítése jóval a mérés előtt megtörtént, ezért az adatbázisokban folytatott keresés kezdeti időpontját 2010. január 1-ben határoztam meg. A keresési időszak záró időpontja 2021. november 30. volt. A

[38] oktatas.hu

mérések dokumentumai esetében a PISA 2009, TIMSS 2019 és PIRLS 2016 mérési ciklusokkal kezdtem a vizsgálatot, tekintet nélkül a dokumentum megjelenési évére.

Az adatbázisokban végzett keresés esetén további feltételként szabtam meg, hogy a találatok lektorált (*peer review*) publikációk legyenek. A mérések saját dokumentumainál ezt a feltételt nem kötöttem ki, mivel ezek nem *peer review* publikációk, de a mérések integritását biztosítandó fontos dokumentumok, melyekre tudományos munkákban is hivatkoznak, illetve saját kutatás tervezésekor is figyelembe vesznek.

A publikációk nyelvét magyar és angol nyelvben határoztam meg. A megszorítás miatt előfordulhat, hogy bizonyos régiókra vonatkozó eredmények kiszorulnak a vizsgálatból, azonban álláspontom szerint a mérések nemzetközi jellege miatt a mérések szervezői által készített dokumentumok és a releváns tudományos eredmények mindenképpen bekerülnek.

A közlemények formai szempontból lehetnek tudományos cikkek, tanulmánykötetben megjelent tanulmányok és könyvfejezetek. Kizárásra kerültek a recenziók, a könyvismertetők, az interjúk, a konferencia-előadások és az absztraktok. További feltétel, hogy a tételek lektoráltak legyenek. Ezt a külföldi adatbázisok esetében a keresési beállításokban szabályoztam, a magyar keresés esetében a megjelenés helyének ellenőrzésével. Erre folyóiratok esetében az MTMT folyóiratkereső felületét alkalmaztam, mely feltünteti a folyóirat tudományos és lektorált jellegét.

A publikációk bevonására tartalmi és módszertani kritériumokat is meghatároztam. Azok a publikációk kerültek bevonásra, melyek empirikus kutatás eredményét vagy ilyen kutatások összegzését közlik. A kutatás minimális mintanagyságát kvantitatív vizsgálat esetében 100 főben, kvalitatív vizsgálat esetében 15 főben határoztam meg. További kritériumként szabtam meg, hogy a kutatás célja a papír-ceruza és számítógépes adatfelvétel közötti, a mérés módjából következő eltérések vizsgálata vagy a két mérési mód összehasonlítása legyen. További megkötés, hogy a kutatás szorosan kapcsolódjon a három vizsgált nemzetközi méréshez, azaz eredeti adatbázisokat vagy eredeti feladatokat használjon fel, illetve a lebonyolítás szorosan kapcsolódjon a felmérésekhez. A két mérési mód vizsgálata kizárja azokat a választható területeket és méréseket, melyek kizárólag számítógépes adatfelvétellel valósultak meg.

6.1.3. Kulcsszavak

A kereséshez a kulcsszavakat két szempont alapján választottam ki. Egyrészt fókuszáltam a kiválasztott három nemzetközi tanulóiteljesítmény-mérésre, melyek mozaikszavai (PISA, TIMSS, PIRLS) megfelelő indikátorok a mérésekhez kapcsolódó összes forrás megtalálásához. A felmérések teljes nevét nem tartottam fontosnak beválasztani, mivel a szövegekben ugyan szükségszerűen megjelenik, de a címekben, az absztraktokban és a kulcsszavak között a rövid változat a jellemző, a teljes szövegben pedig biztosan felbukkan. A hazai adatbázisokban történt keresés esetében nem számítottam feldolgozhatatlan számú találatra, ezért a kulcsszavakat nem bővítettem tovább.

A mérésekhez kapcsolódó teljes nemzetközi szakirodalom várható gazdagsága miatt a nemzetközi adatbázisokban történt keresést további kulcsszavakkal egészítettem ki. Egyrészt a médiahatás angol megfelelőit vontam be, melyek azonban különbözőek az egyes mérések esetében. A magyar médiahatás szó angol eredetije a *„mode effect”*, ami a papír és a számítógépes adatfelvétel eredménye közötti szisztematikus eltérésre utal. A kifejezés általánosan elterjedt, a magyar forrásokban is ez szerepel angol eredetiként (pl. R. Tóth & Hódi, 2011). Ezt a kifejezést használja a PISA (OECD, 2017b, p.152–162), az erre vonatkozó vizsgálatot *„mode study”* kifejezéssel jelöli. Az összefoglalóban (OECD, 2016b) kötőjeles formában jelenik meg *(„mode-effect”*), azonban összesen három alkalommal szerepel. A TIMSS mérés az *„item equivalence”* kifejezést használja (von Davier et al., 2020), ami arra utal, hogy a mérési módok közötti megfeleltetést olyan itemekkel lehet biztosítani, melyek egyformán viselkednek a két adatfelvételi felületen. Az erre vonatkozó vizsgálatot az *„item equivalence study”* kifejezéssel jelöli.

A mérések és a médiahatás kulcsszavaival (*„mode effect”* és/vagy *„item equivalence”*) a nemzetközi adatbázisokban igen kevés, jellemzően 10 alatti találatot kaptam. Ez alapján úgy döntöttem, hogy a témára vonatkozó kulcsszavakat a számítógépes adatfelvételre vonatkozó *„computer-based”* kulcsszóval bővítem, mivel médiahatás-vizsgálat esetében ennek a szónak – a mérések eredeti papír-ceruza (*paper-based*) adatfelvételi módjához képest – biztosan meg kell jelennie. Ez a találatok számának jelentős növekedését hozta. Az egyes adatbázisok, keresőszavak és alkalmazott szűkítő feltételek kombinációját az 1. táblázat tartalmazza.

1. táblázat

Az egyes adatbázisokban futtatott keresések kulcsszavai és beállításai

Adatbázis	Kereső kifejezés	Egyéb kritérium
Arcanum	SZO=(pisa timss pirls) AND DATE=(2010--)	
MATARKA	PISA OR TIMSS OR PIRLS	2010-től
MTMT	PIRLS PISA TIMSS	Jelleg: Tudományos Év: >=2010
EBSCO ERIC JSTOR ProQuest Science Direct Web of Science	(PISA OR TIMSS OR PIRLS) AND ("mode effect" OR "item equivalence" OR "computer-based")	2010-től lektorált

6.1.4. A szakirodalomkeresés folyamata és eredménye

A keresések eredményét a Zotero (zotero.org) hivatkozáskezelő szoftverbe gyűjtöttem, és Microsoft Excel szoftver felhasználásával a metaadatok alapján szelektáltam.

A hazai áttekintés kizárási folyamatát a 12. ábra mutatja be. A találatok egy részében a „PISA” keresőszó a szerző nevében szerepelt, más részében nem magyar vagy angol nyelvű tétel volt a találatok között. Egy tétel latin nyelvűként szerepelt, a szöveg ellenőrzése után magyar nyelvűnek bizonyult, ezért ezen kritérium alapján nem került kizárásra. Két tétel esetében a magyar nyelvű tétellel tartalmilag egyező, azonos szerzőtől származó angol nyelvű publikációt találtam, az angol nyelvű tételeket kizártam. A duplumok és ezen tételek eltávolítása után a tételeket a publikáció címe alapján is ellenőriztem. Az áttekintés során a folyóiratok címe alapján kizártam a nem lektorált tételeket, továbbá formai szempontok alapján a konferenciaanyagokat, recenziókat, könyvismertetőket és interjúkat is. Három tételt online formában nem sikerült megtalálni. Cím és absztrakt alapján 165 tételt zártam ki, ebből 44 tétel a kiválasztott mérésekkel foglalkozik, azonban nem a médiahatás témáját vizsgálja.

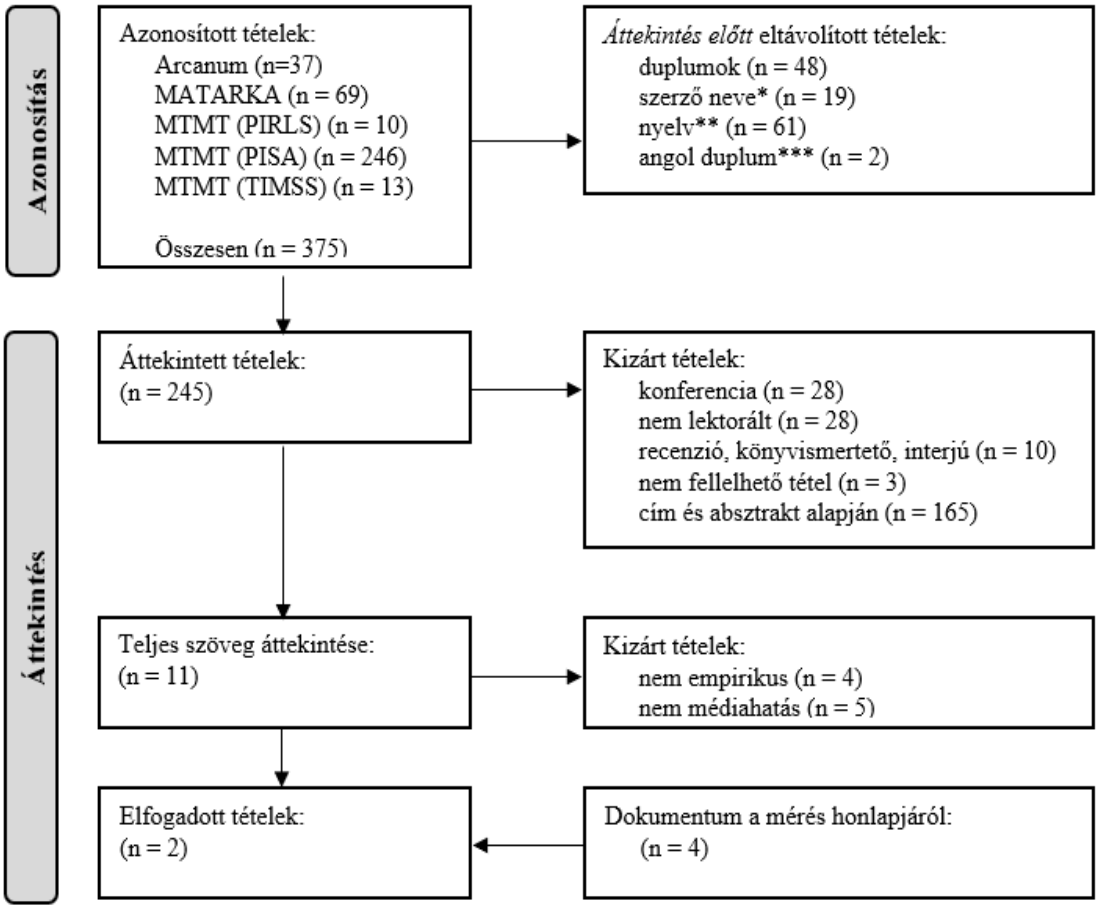
A teljes szöveg áttekintésére 11 publikáció esetében került sor, ezek esetében a téma érinthette a médiahatás kérdését. A tételek harmada nem empirikus megközelítésű volt, a tételek fele pedig nem vizsgálja a médiahatás kérdését. Velkey (2018) munkája foglalkozik az adatfelvételi módok közötti különbséggel és összefoglal néhány

médiahatással foglalkozó eredményt, de nem végez empirikus kutatást a témában, ezért ezt a tételt kizártam az elemzésből. A hivatkozott munkákat áttekintettem további forrásokért, de új forrást nem találtam.

A befoglalási kritériumoknak végül két tétel felelt meg, mindkettő az Oktatási Hivatal mérésekhez kapcsolódó dokumentuma. Az Oktatási Hivatal oldalán további négy, hasonló dokumentumot találtam, melyek a megfelelő mérésekhez kapcsolódnak, és információval szolgálhatnak a kiválasztott három mérés médiahatásával kapcsolatban. Az így kapott hat tétel adatait a 2. táblázat tartalmazza. Mivel ezek mindegyike a mérésekhez kapcsolódó jelentés, ezért a nemzetközi mérési dokumentumokkal együtt dolgoztam fel őket.

12. ábra

A magyar katalógusokban fellelt tételek PRISMA folyamatábrája. Saját ábra (Page et al., 2021) alapján



*: kulcsszó a szerző nevében; **: magyar nyelvű szöveg latinként jelölve; ***: magyar nyelvű szöveggel tartalmilag egyező angol nyelvű szöveg.

2. táblázat

A PIRLS, PISA és TIMSS nemzetközi mérések hazai szervezőjénél (Oktatási Hivatal) fellelt technikai és összegző jelentések listája a megjelenés évének sorrendjében

Forrás	Cím	Adatbázis
Balázsi & Ostorics, 2011	PISA2009 Digitális szövegértés. Olvasás a világhálón	MTMT
Balázsi et al., 2013	PISA 2012 Összefoglaló jelentés	OH
Ostorics et al., 2016	PISA 2015 Összefoglaló jelentés	OH
Balázsi et al., 2017	PIRLS 2016 Összefoglaló jelentés a 4. évfolyamos tanulók eredményeiről	MTMT
Oktatási Hivatal, 2019	PISA 2018 Összefoglaló jelentés	OH
Palincsár et al., 2020	TIMSS 2019 Összefoglaló jelentés	OH

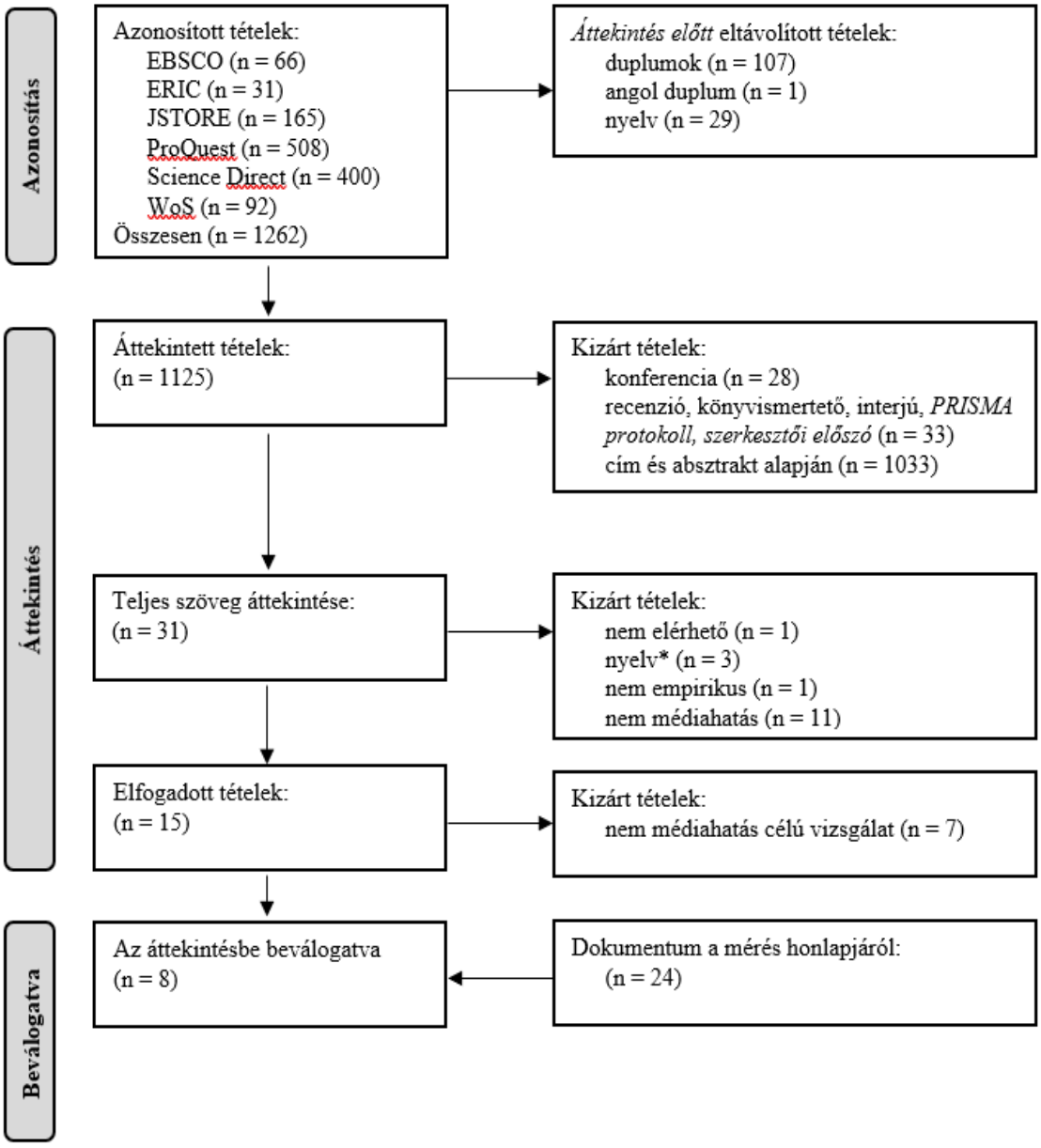
A nemzetközi adatbázisokban talált tételek szelekciós folyamatáról a 13. ábra tájékoztat. Áttekintés előtt a duplumokat és a nem angol nyelvű találatokat zártam ki. A hivatkozáskezelőből exportált nyelvet – ahol eltért az előre meghatározottaktól vagy hiányzott – és a publikáció címét együttesen használtam fel a nyelv szerinti kizárás eldöntésére.

Az áttekintés során 61 tételt zártam ki a publikáció típusa alapján. Itt új típusként megjelent a PRISMA protokoll és a szerkesztői előszó, melyek az adatbázisokban cikként kerültek kategorizálásra. A cím és absztrakt alapján kizárt 1033 publikáció között 107 foglalkozott valamelyik kiválasztott méréssel, de témája nem a médiahatás vizsgálata volt. Ide tartoznak a trendvizsgálatok, a PISA csak számítógépes adatfelvétellel mért választható területeivel, jellemzően a komplex problémamegoldás (*Complex Problem Solving*) vagy a kollaboratív problémamegoldás (*Collaborative Problem Solving*) mérésével kapcsolatos vizsgálatok, a számítógépes adatfelvétel naplófájljainak vizsgálatai, a tanulói kérdőív IKT eszközök használatával kapcsolatos válaszainak kutatása. Ezen a ponton még nem zártam ki azokat a publikációkat, melyek a PISA 2009 és 2012 választható digitális szövegértését vizsgálták, feltéve, hogy a főmérés szövegértés területével vetették össze. 27 tétel a médiahatás vizsgálatára irányult, de nem valamelyik kiválasztott mérés adatait vagy keretét használta fel. A kódolás minőségének ellenőrzésére a tételek közel 20%-át (221 tétel) másodkódolásnak vetettem alá. Mindösszesen 5 tétel esetében különbözött a kódolók véleménye, a Cohen-kappa (Cohen

κ = 0,69) alapján ez jelentős egyezést jelent (Landis & Koch, 1977). Az eltérően kódolt tételeket a kódolók a teljes szöveg áttekintése alapján kizárták.

13. ábra

A nemzetközi adatbázisokban és a mérések dokumentumai között fellelt tételek PRISMA folyamatábrája. Saját ábra (Page et al., 2021) alapján



*: angol cím és absztrakt, teljes szöveg nyelve alapján kizárt.

A tételek teljes szövegének vizsgálatára 31 tételt fogadtam el. A teljes szöveg egy esetben nem volt elérhető. A tételek közül három angol címmel és absztrakttal szerepelt, de a teljes szöveg idegen nyelvűnek bizonyult (bolgár, koreai és német), így a nyelv alapján összesen 32 (áttekintés előtt 29, teljes szöveg áttekintésekor további 3) tétel került kizárásra. A feldolgozott 27 publikációból egy nem empirikus munka, további 11 kutatás célja pedig nem a médiahatás vizsgálatára irányult.

A fennmaradt 15 publikáció két markáns csoportba volt osztható. Az egyikbe olyan publikációk tartoznak, melyek kifejezetten a kutatási kérdésben meghatározott médiahatást vizsgálják, azaz az itemek papír-ceruza és számítógépes adatfelvételből származó jellemzőinek összehasonlítására irányulnak. A másik csoport olyan tételekből áll, melyek elsősorban a lineáris és nemlineáris (jellemzően online vagy digitális) szövegek olvasása közötti különbségeket vizsgálják különböző háttértényezők (például IKT használat) szerint, és a kétféle szövegértési folyamatot a kétféle adatfelvétellel mérik. Jellemzően a PISA 2009 és 2012 digitális szövegértés részterületen és a szövegértés főterületen elért eredményt, illetve a 2016. évi PIRLS és ePIRLS mérések eredményét vetik össze. A szövegek kódolását két kódoló végezte, két tétel esetében volt eltérő vélemény (kizárandó vagy nem médiahatás célú kapcsolatvizsgálat). A Cohen-kappa (Cohen $\kappa = 0{,}89$) alapján ez majdnem tökéletes egyezés (Landis & Koch, 1977). Mivel a két kategória egyike sem került bele a végső feldolgozásba, ezért a vélelmes tételek esetében nem volt szükséges egyezségre jutni. Az elemzésbe beválasztott nyolc publikáció adatait a 3. táblázat tartalmazza.

3. táblázat

A nemzetközi adatbázisokban folytatott keresés eredménye

Forrás	Cím	Kiadvány	Adatbázis
Fishbein et al., 2018	The TIMSS 2019 Item Equivalence Study: Examining Mode Effects for Computer-Based Assessment and Implications for Measuring Trends	*Large-Scale Assessments in Education, 6*	ERIC, ProQuest, WoS
Hamhuis et al., 2020	Tablet assessment in primary education: Are there performance differences between TIMSS' paper-and-pencil test and tablet test among Dutch grade-four students?	*British Journal of Educational Technology*, *51*(6), 2340–2358	EBSCO, ERIC, WoS
Jerrim, 2016	PISA 2012: how do results for the paper and computer tests compare?	*Assessment in Education: Principles, Policy & Practice*, *23*(4), 495–518	EBSCO, ERIC, WoS
Jerrim et al., 2018	PISA 2015: how big is the 'mode effect' and what has been done about it?	*Oxford Review of Education, 44*(4), 476–493	EBSCO, ERIC, WoS
Kroehne et al., 2019	Construct Equivalence of PISA Reading Comprehension Measured With Paper-Based and Computer-Based Assessments	*Educational Measurement: Issues & Practice*, *38*(3), 97–111	EBSCO, WoS
Robitzsch et al., 2020	Reanalysis of the German PISA Data: A Comparison of Different Approaches for Trend Estimation With a Particular Emphasis on Mode Effects	*Frontiers in Psychology* (Vol. 11)	WoS
Zehner et al., 2019	Unattended consequences: how text responses alter alongside PISA's mode change from 2012 to 2015	*Education Inquiry*, *10*(1), 34–55	EBSCO, ERIC, ProQuest, WoS
Zehner et al., 2020	PISA reading: Mode effects unveiled in short text responses	*Psychological Test and Assessment Modeling*, *62*(1), 85–105	ProQuest

A nemzetközi mérések szervezőinek honlapjáról összesen 24 dokumentumot gyűjtöttem össze a mérés éve és a kötet címe alapján. A dokumentumok összegzett jellemzőit a 4. táblázat tartalmazza. Ehhez a 24 forráshoz vettem hozzá a hazai szervező honlapján talált hat dokumentumot. A mérési dokumentumok jellemzően a mérések keretrendszerét, a mérés eredményeit és a technikai részleteket közlik. A három különböző célú kiadvány kiegyensúlyozottan szerepel a gyűjtésben, azonban a mérések közül a PISA nagyobb arányt képvisel. Ennek oka, hogy míg a TIMSS és PIRLS mérések lényegében egy-egy számítógépes ciklussal szerepeltek a keresésben meghatározott időintervallumban, addig a PISA esetében két számítógépes kiegészítő és két teljesen számítógépes mérés szerepelt.

4. táblázat

A nemzetközi mérések saját dokumentumainak összegzett jellemzői. Az egyes cellákban a mérések adott célú dokumentumainak száma található

Mérés	Mérési keret	Mérés eredményei	Technikai jellemzők	Egyéb	Összesen	Hazai dokumentum
PIRLS	2	2	1	–	5	1
TIMSS	2	1	1	–	4	1
PISA	4	5	4	2	15	4
Összesen	8	8	6	2	24	6

A szövegértés és a digitális szövegértés eredményeket háttértényezők mentén összehasonlító publikációkat a teljes szöveg áttekintése során kizártam jelen elemzésből. A matematikai műveltség papír-ceruza és számítógépes eredményeinek vizsgálatával foglalkozó forrásokról – a mérési keret azonossága miatt – a kutatás célja szerint döntöttem. Amennyiben a digitális/számítógépes készségek és egyéb háttértényezők kapcsolatainak vizsgálatára irányult, akkor kizártam, amennyiben a tartalmi keret egyezésén alapult, és a médiahatás vizsgálatára irányult, akkor belefoglaltam az elemzésbe.

Nem találtam olyan publikációt, mely ugyanabban a vizsgálatban több nemzetközi mérés adatait is felhasználja, ezért az eredményeket a mérések szerinti tagolásban mutatom be, összevonva a hazai és nemzetközi mérési dokumentumok és a nemzetközi adatbázisok találatainak eredményeit. A nemzetközi adatbázisokban és a mérési dokumentumok között nem találtam a PIRLS méréshez szorosan kapcsolódó,

médiahatás vizsgálatára irányuló kutatást vagy leírást. A lehetséges publikációk jellemzően a lineáris és nemlineáris (internetes) szövegértés tanulói eredményeit vizsgálják különböző háttértényezők (nem vagy IKT használat) mentén. Ezeket a kutatásokat a mérések keretbeli különbségei alapján nem tekintettem médiahatás vizsgálatnak és kizártam a további elemzésből.

A hazai mérési dokumentumok egy kivétellel a méréshez kapcsolódó összefoglaló jelentések, a hatodik kifejezetten az egyik részterület, a PISA 2009 digitális szövegértés eredményeivel foglalkozik. Az egyes jelentések jellemzően szűkszavúan, egy-egy bekezdés erejéig foglalkoznak a médiahatás kérdésével. A nemzetközi adatbázisokból leválogatott 8 tétel esetében csak azokat a kutatási kérdéseket és eredményeket tárgyaljuk, amelyek a médiahatás vizsgálatával foglalkoznak.

6.1.5. PISA

A 2015. évi PISA mérés az ezt vállaló országokban teljes egészében számítógépes formában valósult meg (OECD, 2016b). A hazai szervező az új mérési móddal kapcsolatban arról tájékoztat, hogy a médiahatást a próbamérés során vizsgálták, és nemzetközileg egységesen a trendek kiszámításakor figyelembe vették (Ostorics et al., 2016). A 2018-as mérés összefoglalója (Oktatási Hivatal, 2019) a 2015 előtti és utáni eredmények összehasonlításánál óvatosságra int, valamint hivatkozik egy hazai elemzésre (Lak, 2020), amelynek célja a médiahatás magyarországi vizsgálata, azonban ez a publikáció nem lektorált tudományos folyóiratban került megjelentetésre, ezért elemzésembe nem vontam be.

A mérési keret (OECD, 2017a) a médiahatással kapcsolatban említ korábbi vizsgálatokat és azok ellentmondásos eredményeit, valamint megemlíti az OECD PIAAC médiahatással kapcsolatos eredményét is. A dokumentum szerint a médiahatás vizsgálatát a 2015-ös mérés próbamérése során végezték, a leírás külön forrásként található meg (OECD, 2016a), és itt található a médiahatás első PISA definíciója. E szerint „a médiahatás (*mode effect*) kifejezés arra a megfigyelésre utal, hogy az egyik módban (például papír alapon) bemutatott feladatok másképp működhetnek, mint amikor egy másik (pl. számítógépes) módban mutatják be őket” (OECD, 2016a, 4). A trend itemek vizsgálata klasszikus és modern tesztelméleti (IRT) modellek segítségével történt. Az itemparaméterek korrelációja mind a nehézség, mind a meredekség paraméterek esetében 0,9 feletti. A médiahatást becslő IRT modellek összehasonlítása alapján nem szükséges

országonként vagy személyenként, elegendő itemenként alkalmazni a paraméterek eltolását. Ez egyrészt jelenti a mért konstruktumok azonosságát, másrészt az itemek paramétereinek egyedi vizsgálatát. A trend itemek nagy része skálainvariánsnak bizonyult, azaz mindkét paramétere megegyezik a két módban, ami lehetővé teszi a két mód összekötését. A metrikusan invariáns itemek, azaz amelyek esetében a nehézség különbözött, mindkét irányú médiahatást mutattak. Az ország-mód és a nem-mód kereszthatást regressziós modellekkel tesztelték, szignifikáns ($p < 0,05$) kereszthatást nem találtak, azaz az itemenként végzett korrekció továbbra is biztosítja az egyes nemek és országok közti összehasonlíthatóságot. Felhívják ugyanakkor a figyelmet a kis mintanagyságra (országonként kb. 400 papír-ceruza és 600 számítógépes tesztkitöltés), mint az eredmények érvényességének korlátjára.

A főmérés technikai leírása (OECD, 2017b) szerint az itemeket a próbamérés eredményei alapján elemezték. Szintén bemutatják a próbamérés során végzett vizsgálatot. A főmérés trend itemeinek nagy része invariánsnak bizonyult, az összes item 90%-a legalább metrikusan invariáns volt (5. táblázat). Az itemek paramétereinek eltolása és az invariáns itemek segítségével a 2015. évi mérés eredményeit a korábbi trendhez kötötték. Itt érdemes megemlíteni, hogy a 2015. évi mérés során a skálák számításának módját két további ponton, az IRT modell megválasztásában és a korábbi ciklusok skáláinak felhasználásában is megváltoztatták. A PISA 2018 dokumentumai nem tartalmaznak médiahatásra vonatkozó új információt, jellemzően a 2015-ös mérés dokumentumaiban közölt információkat közlik.

5. táblázat

A közös és egyedi paraméterezésű itemek százalékos aránya a PISA 2015 egyes mérési területein (Forrás: OECD, 2017b. p.225 alapján)

Itemek megoszlása	Matematika	Szövegértés	Természettudomány
% egyedi paraméter (csoportra jellemző)	2,16%	3,01%	2,62%
% egyedi paraméter (néhány csoportra jellemző)	3,36%	7,98%	7,68%
% metrikusan invariáns közös/nemzetközi paraméterek	33,22%	30,33%	20,96%
% skála invariáns közös/nemzetközi paraméterek	61,25%	58,68%	68,74%
Mód és az itemek száma a PISA 2015 főmérésben	PBA item: 83 CBA item: 81	PBA item: 103 CBA item: 103	PBA item: 85 CBA item: 85

PBA: papír-ceruza teszt (paper based assessment), CBA: számítógépes teszt (computer based test)

A nemzetközi adatbázisban történt keresés eredményeként kapott 8 találat közül 6 tétel (Jerrim, 2016; Jerrim et al., 2018; Kroehne et al., 2019; Robitzsch et al., 2020; Zehner et al., 2019, 2020) foglalkozik a PISA méréshez kapcsolódó médiahatás vizsgálatokkal. Ezek körében is két jól elkülöníthető csoport alakítható ki. A Jerrim (2016; 2018), Kroehne és munkatársai (2019), valamint Robitzsch és munkatársai (2020) nevével jelzett források a próbaméréshez hasonló vizsgálatokat végeznek, a médiahatást az itemek jellemzőinek és az egyes területeken mért teljesítményeknek az összehasonlításával elemzik. Zehner és munkatársai (2019, 2020) egymásra épülő kutatásokat mutatnak be, melyek a nyílt válaszok bizonyos jellemzőinek a két tesztmédium közötti különbözőségét vizsgálják. Az egyes források módszertani jellemzőit és fő eredményét a tartalmazza. A mérési adatok forrása négy esetben kizárólag Németország. Az érintett mérésekben megjelennek 2012. és 2015. évi főmérések, a 2015-ös mérés próbamérése, valamint egy, a 2012. évi főméréshez kapcsolódó német kiegészítő vizsgálat. Ennek során a 2009-es ciklus 35 papír-ceruza alapú szövegértés itemét számítógépre adaptálták, és egy almintán felvették a mérés másnapján. A vizsgálatokban használt elemzési módszerek magas minőségűek, a mintaelemszámok szintén megfelelőek.

6. táblázat

A nemzetközi adatbázisban történt keresés eredményeként kapott kutatások módszertani jellemzői és fő eredménye

Forrás	Mérés	Terület	Mintanagyság	Módszer	Eredmény
Jerrim, 2016	2012 matematika és szg. matematika	32 gazdasági egység	~200 000 fő	Ország szintű korreláció, LPM	Eltérés az országok harmadánál (10–20 pont)
Jerrim et al., 2018	2015 próbamérés	Németország, Svédország, Írország	3438 fő	OLS regresszió	Kis mértékű, területenként különböző eltérés (10–20 pont)
Kroehne et al., 2019	2012* szövegértés	Németország	856 fő	SEM	Egyező konstruktum, Kis, egyenetlen médiahatás
Robitzsch et al., 2020	2015 próbamérés	Németország	1023 fő	IRT, Jackknife szimuláció	Kis mértékű, területenként eltérő médiahatás (10–20 pont)
Zehner et al., 2019	2012, 2015 szövegértés	Németország	43 396 válasz	GLMM	Hosszabb és több információt tartalmazó válaszok
Zehner et al., 2020	2012* szövegértés	Németország	7495 válasz	GLMM	Hasonló eredmény, kisebb különbség

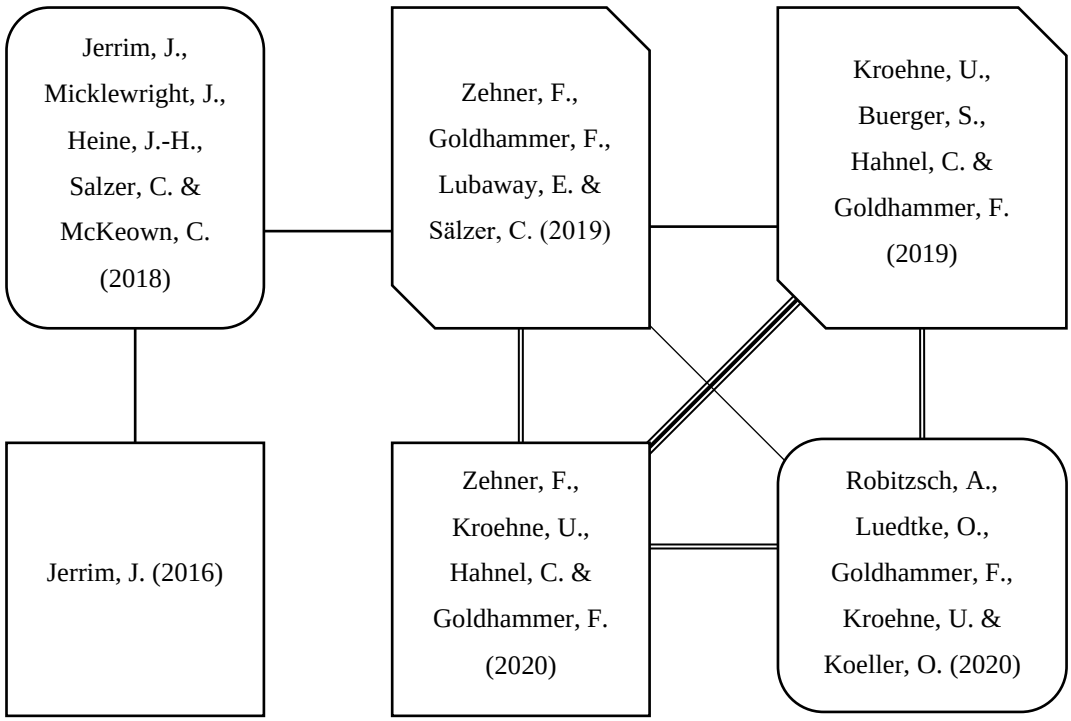
*: A PISA 2012 méréshez kapcsolt német médiahatás-vizsgálat, PISA 2009-es itemek papír-ceruza és számítógépes változatával.

LPM = lineáris valószínűségi (linear probability) modell regresszió; OLS = lineáris regresszió legkisebb négyzetek módszerével (ordinary least squares); SEM = strukturális egyenletek módszere (structural equation modelling); IRT = modern tesztelméleti modell (item response theory); GLMM = általánosított lineáris vegyes modell (general linear mixed-model) regresszió

A források közötti kapcsolatot a közös szerzőkkel jelezve (14. ábra) kirajzolódik a PISA méréshez kapcsolódó források mögötti szakmai együttműködés hálózata, melynek magját Goldhammer és Kroehne adják 4, illetve 3 szerzőséggel.

14. ábra

A PISA méréshez kapcsolódó médiahatás-vizsgálatok kapcsolatai a közös szerzők alapján. (Forrás: saját ábra)



Megjegyzés. A kapcsolati vonalak száma a közös szerzők számára, az alakzatok formája a közös adatforrásra utal.

A PISA méréshez kapcsolódó források mindegyike hasonló megállapításokra jut. A papír-ceruza és számítógépen mért konstruktumok nem különböznek egymástól, legalábbis nem olyan mértékben, ami veszélyeztetné a trendek folytonosságát. Ezt egyrészt az itemparaméterek, másrészt a kétféle adatfelvételben elért képességpontok magas korrelációja alapján állítják, amit Kroehne és munkatársai (2019) strukturális egyenletek módszerével is ellenőriz. Az itemek kis része esetén figyelhető meg a médiahatás miatti eltérés a paraméterekben. 35 feladatból öt nehezebb, egy könnyebb a számítógépes mérésben, míg a meredekségben nem volt különbség a mérési módok

között. A teljesítmény szerinti eltérés szignifikáns, a két tesztfelvételi mód között jellemzően 10 és 20 pont között mozog, és a próbamérésen végzett elemzés alapján nem egyformán érinti az egyes mérési területeket. Ekkora eltérés az évek közötti összekötés hibája 4–6-szorosának (OECD, 2014a, p.281), az iskolai évfolyamok közötti különbség (41 pont) 25–50%-ának, illetve a Magyarország 2012-es matematika eredménye (477 pont) és az OECD átlag (494 pont) közötti szignifikáns különbségnek feleltethető meg (OECD, 2014a, p.46–47). Ezek az eredmények megfelelnek a mérés saját dokumentumaiban szereplő információknak, ugyanakkor alátámasztják, hogy az egyes országokban különböző mértékű médiahatásra lehet számítani, amit a trendszámítás nem vesznek figyelembe.

A nyílt válaszok elemzései (Zehner et al., 2019, 2020) nem egyformán kapcsolódnak a médiahatáshoz. A 2019-es publikáció a PISA 2012 és 2015 közös szövegértés itemeire adott szöveges válaszokat hasonlította össze, így az eredmény a populációkból származó különbséget is hordozhatott. A cikk ennek megfelelően médiahatás helyett ciklushatásnak nevezi a jelenséget. A 2020-as publikáció ugyanezzel a kutatási kérdéssel foglalkozik, kifejezetten a médiahatásra fókuszálva a 2012-es kiegészítő mérés nyílt válaszain, és a korábbihoz hasonló, bár kevésbé markáns eredményt kaptak. A nyílt válaszok értékelésének szempontjai a megjelenő elemek száma és az információtartalom voltak. A vizsgálatok a számítógépes adatfelvétel válaszaiban több válasz-elemet és valamivel több információt találtak.

6.1.6. TIMSS

A TIMSS 2019 mérés hazai jelentése (Palincsár et al., 2020) a bevezető fejezetben tájékoztat a számítógépes adatfelvételről. Ez alapján a papír-ceruza és számítógépes tesztek tartalomban és felépítésben egyeztek (nagyrészt a papír-ceruza TIMSS 2015 itemeit vették alapul), valamint a nemzetközi eredmények és a trend összehasonlíthatósága érdekében a számítógépes mellett hozzávetőlegesen feleakkora mintán papír-ceruza adatfelvétel is történt. A trendeket két lépésben kapcsolták össze: először a hagyományos adatfelvételt a 2015. évi méréshez, majd a számítógépes adatfelvételt a 2019. évi papír-ceruza mérés eredményeihez. Ez az információ minden tekintetben egyezik a mérés eredményeinek nemzetközi dokumentumával (Mullis et al., 2020), amely szintén nem ad további technikai részleteket. Ugyanitt az eredményeket

közlő táblázatokban külön-külön szerepelnek az eTIMSS és a paperTIMSS országok eredményei.

A mérési keret dokumentuma (Mullis & Martin, 2017) tájékoztat az egyes mérések tesztfüzeteinek összeállításáról. Ez alapján a méréshez 14 (az innovatív, például húzd-és-vidd (*drag-n-drop*) itemektől eltekintve) teljesen egyező tesztfüzetet alakítottak ki, azaz a TIMSS 2019 mérést úgy tervezték meg, hogy 1) a mérőeszközök csak a közvetítő médiumban különböznek, ezáltal alkalmasak a médiahatás mérésére, és 2) a trendek és a két mérési mód megbízható összehasonlíthatósága érdekében ezt a médiahatást figyelembe is vették az eredmények közlésekor.

A médiahatás számítását és a skálák összekötésének módját a technikai leírás (Martin et al., 2020) ismerteti. A dokumentum szerint a tartalmi validitás és a médiahatás vizsgálatára 2017-ben, a próbamérést megelőző évben került sor az eTIMSS *Item Equivalence Study* keretében. A vizsgálat további célja volt az itemek digitális változatának operatív tesztelése és ezzel a próbamérés előkészítése. Az eredményeket tudományos cikk formájában jelentették meg (Fishbein et al., 2018), ami éppen a nemzetközi adatbázisban történt keresés egyik találata. A technikai leírás nem közöl részletes eredményeket, hanem a cikkre hivatkozik.

Fishbein és munkatársai (2018) vizsgálata 25 országban, két mérési területen, két évfolyamon zajlott 26 000 tanuló bevonásával. Magyarország nem vett részt a vizsgálatban. A mérés konstruktumát (matematika, illetve természettudomány) egyezőnek találták a két mérési módban. A híd itemek 80%-át invariánsnak, azaz a két mérési módon megegyező mérési tulajdonságúnak találták. Az eltérő itemek jellemzően valamilyen technikai probléma vagy innovatív eszköz miatt lehettek különbözőek. Az invariáns itemek százalékos megoldottságai összességében kismértékű médiahatást mutattak, a papír-ceruza mérést könnyebbnek jelezve. Meredekségben és a hiányzó/el nem ért válaszok arányában nem volt különbség. A két mérési mód alapján számított teljesítményben matematikából 14, természettudományból 7 pontnyi átlagos különbség rajzolódott ki, ami szignifikáns és összemérhető az elfogadható mértékű mérési hibával (annak 1–2-szerese), vagy az évfolyamok közötti különbséggel (ami az általános iskolai évfolyamok közötti különbség negyede), vagy a Magyarország 2019-es matematika eredménye (523 pont) és a TIMSS középérték (500 pont) különbségének felével. Ez indokolja a főmérés során a számítógépes adatfelvételt kiegészítő papír-ceruza adatfelvételt és a médiahatás figyelembevételét a skálák igazításakor. Háttérjellemzők szerinti különbséget nem találtak.

A TIMSS méréshez kapcsolódó másik találat Hamhuis és munkatársainak (2020) az *Item Equivalence Study*-ra irányuló kritikai vizsgálata. Az eredeti adatfelvétel holland adatainak másodelemzését az motiválta, hogy a holland általános iskolákban elterjedt a tablethasználat, ami a számítógép mellett az eTIMSS másik elfogadott adatfelvételi médiuma. A mintát 25 iskolából 532 tanuló alkotta, mindegyikük 4. évfolyamos. A két módban elért eredmények között nem találtak szignifikáns különbséget ($p > 0,05$), ami felveti a holland TIMSS trend felülbecslését a 2019-es felmérésben. Emellett az eredeti tanulmánytól eltérően kis különbséget találtak a nemek között.

A 2019-es főmérés adatain végzett, médiahatás-vizsgálattal kapcsolatos elemzéseket a technikai leírás (Martin et al., 2020) tartalmazza. A 10–12. fejezetek részletesen bemutatják az itemek ellenőrzésének módját és a skálázási eljárást, amit a kiegészítő papír-ceruza adatfelvétel tesz lehetővé. Az eTIMSS méréshez képest feleakkora mintán a TIMSS 2015 trend-itemeinek papír-ceruza alapú kitöltése is megtörtént. A trend-itemek több, mint 80%-a bizonyult invariánsnak (Martin et al., 2020, p.12.54), ezek adták a skálázás alapját. A 13. fejezet az egyes országokra számított médiahatást és annak nagyságát vizsgálja. Az invariáns itemekkel számított százalékos eredmények alapján kis számú országban található szignifikáns ($p < 0,05$) eltérés, ami nem jelenik meg mindkét mérési területen. Magyarország esetében 3 és 7 pont közötti az eltérés a papír-ceruza és a számítógépes adatfelvétel eredménye között, ami egyik területen vagy évfolyamon sem jelent szignifikáns eltérést.

6.1.7. Összegzés

Magyarország jelenleg három nemzetközi tanulóiteljesítmény-mérésen vesz részt. A PISA, TIMSS és PIRLS mérések különböző korosztályokat mérnek szövegértés, matematika és természettudomány területeken. Mindhárom mérés az utóbbi mérési ciklus(ok) során papír-ceruza mérési módról számítógépes adatfelvételre váltott, amit több tényező indokolt. Az érvek között szerepel a 21. századi képességek bevonása a mérésbe (OECD, 2017a), az innovatív, a terület tartalmi elemeit jobban mérő itemek lehetősége (Mullis & Martin, 2017) vagy a különböző szintű mérések kombinálásának lehetősége (Mullis & Martin, 2019). A felmérés módjának váltása, a trendek további számíthatóságának és összehasonlíthatóságának érdekében, szükséges a különböző mérési ciklusokban alkalmazott mérőeszközök azonossága, vagy legalábbis a különbségekből származó lehetséges eltérések pontos kontrollálása.

A fentiekhez módszertanában hasonló, három évfolyamon teljes körű hazai mérés, az OKM a 2022. évtől szintén számítógépes formában valósult meg (Oktatási Hivatal, 2021). Az egymást követő évek eredményének összehasonlítása csak akkor lehetséges, ha a mérési mód cseréje esetén is azonos skálára kerülnek az egyes eredmények. A mérés médiumának cseréje kapcsán kutatásom céljaként tűztem ki a nemzetközi tanulóiteljesítmény-mérések hazai és nemzetközi dokumentumainak áttekintését, valamint a papír-ceruza tesztelésről a számítógépes tesztelésre való áttérés következményeivel foglalkozó elemzések, azaz a médiahatás-vizsgálatok összegyűjtését és elemzését.

A mérési dokumentumok alapján a digitális szövegértés mérési területek (PISA 2009 és 2012, ePIRLS 2016 és 2021) nem tekinthetők a papír-ceruza szövegértés tesztek számítógépes változatának, mivel ezekben eltérő típusú vagy célú szövegek feldolgozását követelik a tanulóktól (Mullis & Martin, 2015; OECD, 2009). A matematikai műveltség számítógépes mérése (PISA 2012) a papír-ceruza mérés tartalmi keretében került értelmezésre, ezért amennyiben a kutatás célja és módszertana megfelelő, megszorításokkal alkalmas lehet a médiahatás vizsgálatára.

A médiahatás-vizsgálat a PISA mérés esetében a 2015. évi mérés próbamérésének keretében (OECD, 2016a, 2017b), a TIMSS 2019 esetében pedig önálló vizsgálattal történt (Fishbein et al., 2018). Utóbbi bemutatása lektorált tudományos folyóiratban cikként jelent meg. A két vizsgálat közös eleme, hogy a teljes mérésre koncentrál, fókuszában 1) a két mérés konstruktum-azonosságának vizsgálata, 2) az itemek szintjén történő médiahatás-vizsgálat, és 3) a teljesítmények szintjén történő médiahatás-vizsgálat, egyszersmind a trendek folytonosságának biztosítása áll. Mindkét vizsgálat megerősíti, hogy a két tesztelési módban mért konstruktum azonosnak tekinthető, csak az itemek kis részében adódik szignifikáns különbség az itemparaméterekben. Míg a PISA mérés a két módban nem invariáns itemek esetében mindkét irányú médiahatásról beszámol, addig a TIMSS vizsgálat jellemzően a számítógépes adatfelvételt találta nehezebbnek. Az itemek meredekségében nem mutattak ki médiahatást, azaz az itemek viselkedése azonos a két módban. Mindkét mérés esetében szükségesnek találták valamilyen korrekció alkalmazását, azonban az itemek kis részénél határoztak meg országspecifikus item paramétereket. A médiahatás összességében 10–20 pontnyi eltérést mutat az egyes területeken (és évfolyamokon), amiért a trendek számításában érvényesítik a médiahatás korrekcióját. Ekkora különbség körülbelül 6–8 helyezésnek felel meg a mérések rangsoraiban (a középértékek közelében). A PISA esetében a

próbamérés eredménye alapján alkalmazott nehézség paraméterrel, a TIMSS esetében a méréssel együtt felvett papír-ceruza kiegészítő méréssel számították ki a szükséges korrekciókat.

A mérések hazai szervezője dokumentumaiban tájékoztat a módszertani változásokról, és felhívja a figyelmet az ebből származó bizonytalanságra, önálló vizsgálatot lektorált folyóiratban nem publikált. A hazai tudományos irodalom áttekintése során számos publikációt találtam mind a mérésekkel (lásd az Educatio 2015/2 tematikus száma), mind a médiahatás vizsgálatával kapcsolatban (az eDia-hoz kapcsolódóan Herczegné Goldschmidt, 2016; R. Tóth & Hódi, 2011), vagy általában a médiahatás részeként a konstruktum-validitásról (szintén az eDia kapcsán Hülber, 2012). Olyan forrást, mely a médiahatást a három nemzetközi mérés kontextusában vizsgálja – a fent említett jelentéseken kívül – nem találtam.

A nemzetközi adatbázisokban végzett szisztematikus keresés során nyolc publikációt találtam, melyek a médiahatást valamely mérés adatbázisai vagy feladatai segítségével vizsgálja. A PIRLS méréshez nem kapcsolódik elemzés, aminek elsődleges oka lehet, hogy az első számítógépes megvalósítás 2021-ben zajlott[39]. A TIMSS esetében a korábban említett vizsgálaton kívül egy, a PISA-hoz kapcsolódóan hat kutatást találtam. Ezek jellemzően abból a szempontból vizsgálják a kérdést, hogy a mérésekben alkalmazott egységes skála-korrekció megfelel-e az egyes országok esetében. Az eredmények, akárcsak a médiahatás vizsgálatának általános eredményei, eltérő képet mutatnak. Úgy tűnik, hogy a médiahatás bizonyos országoknál negatív, másoknál pozitív irányú lehet a papír-ceruza mérés eredményeihez képest (Jerrim, 2016), de az is előfordul, hogy nincs kimutatható médiahatás (Hamhuis et al., 2020). Ennek következménye lehet, hogy az egyes országok trendje alá- vagy felülbecsli a valóságos eredményt, bár ez a torzítás jellemzően nem jelentős. A médiahatás Zehner és munkatársai (2019, 2020) kutatása alapján kimutatható a számítógépes adatfelvétel szöveges válaszainak nagyobb elem-gazdagságában és információtartalmában.

A feldolgozott mérési dokumentumok és cikkek alapján úgy vélem, hogy az OKM esetében a mérési felület cseréjét előkészítő vizsgálatok (Molnár et al., 2015) után is szükség lehet a médiahatás empirikus feltérképezésére, alkalmasint annak figyelembevételére a 2022-es eredmények kiszámításában. Ehhez leginkább a számítógépes adatfelvételhez illesztett papír-ceruza kiegészítő teszt a legalkalmasabb. A

[39] A források gyűjtése 2021. december 2-án zárult.

nemzetközi mérések eredményei alapján megfontolandó mérési területenként és évfolyamonként külön-külön végezni a vizsgálatot és a korrekciót. Várhatóan az itemek viselkedése nem, csak nehézsége különbözik a két mérési módban, és a papír-ceruza teszttel megegyező megjelenésű számítógépes itemek jelentős része invariáns lesz. Az OKM esetében is javasolható a trendek folytatólagos eredményeinek vagy a tanulók különböző mérési módban felvett adatainak óvatos értelmezése, ahogy – elsősorban a PISA esetében – a nemzetközi mérések hazai szervezője és a módszertani változásokat tárgyaló publikációk jelzik.

6.1.8. Korlátok és kitekintés

A kutatásnak számos korlátja van, melyek egyrészt a szisztematikus áttekintéssel kapcsolatosak, másrészt a fellelt szakirodalmak limitációiból származnak. A hazai áttekintés korlátja, hogy nincs kifejezetten erre a célra kialakított adatbázis, ezért lehetséges, hogy a keresés során kimaradtak olyan publikációk, melyek címében nem szerepel egyik mérés mozaikszava sem, vagy nem szerepel a MATARKA adatbázisban. Ez utóbbi hiányosságot az Arcanum adatbázisában folytatott szövegszerű kereséssel küszöböltem ki, de új forrást nem találtam.

A nemzetközi adatbázisokban folytatott keresés során olyan adatbázisokban folytattam keresést, melyekhez intézményi hozzáférésem volt. Ezek mindegyike elfogadott, megbízható forrás a neveléstudomány területén, de nem teljes körű. További korlát, hogy kizárólag angol vagy magyar nyelvű forrásokat fogadtam el, ami bizonyos régiókban folytatott vagy publikált kutatások kizárását jelenthette, ugyanakkor a teljes szövegek áttekintése során találkoztam távol-keleti, kelet-európai, dél-afrikai és dél-amerikai adatokon végzett vagy ottani szerzőségű publikációkkal. A befoglalási és kizárási kritériumok alapján az úgynevezett szürke irodalomba (*grey literature*) tartozó források (Dobó, 2000) nem kerültek bele a válogatásba. Ennek részben kutatási kapacitással kapcsolatos okai vannak, másrészt igyekeztem magas minőségű, lektorált, tudományos folyóiratban megjelent forrásokat felkutatni.

Az elemzésbe bevont források a PISA esetében jellemzően németországi adatok feldolgozását jelentette, a mérési dokumentumokon felül vizsgált nyolc tanulmány majd mindegyike európai kontextusban vizsgálódik. A távol-keleti és dél-amerikai szempontok egy publikációban jelentek meg (Jerrim, 2016). A különböző mérések adatain folytatott összehasonlítások esetében nem zárható ki a minták populációból adódó különbsége, ami

további bizonytalanságot eredményez a médiahatás meghatározásában. További probléma a PISA mérés esetében, hogy a számítógépes mérési mód bevezetésével egy időben más módszertani változások is történtek. Egyrészt az itemparaméterek paraméterezése (egyparaméteres helyett kétparaméteres modell illesztése), másrészt a skálák kialakítása (az itemparaméterek kialakításakor minden korábbi mérési ciklust figyelembe vettek), harmadrészt az el nem ért itemek kezelése (helytelen válasz helyett az el nem ért itemek nem számítottak bele a pontszámításba) is megváltozott. Ezek egy része szintén vizsgálat tárgyát képezte (Robitzsch et al., 2020), azonban hatással lehetnek a médiahatás meghatározására is.

A nemzetközi tanulóiteljesítmény-mérések tapasztalatai arra engednek következtetni, hogy a papír-ceruza tesztek itemei – amennyiben a lehető leghasonlóbb módon kerülnek átültetésre – igen nagy arányban hasonló viselkedést mutatnak számítógépes teszt esetében is. Ennek alapján arra lehet következtetni, hogy a papír-ceruza teszteken bemért itemparaméterek jól modellezik ugyanezen itemek számítógépes változatainak paramétereit, értéküket szimulációhoz (ld. 5.1 és 6.4 fejezet) felhasználva érvényes eredményeket kapunk.

6.2. Lineáristól az adaptív mérés felé – a nyílt itemek szerepe[40]

Az adaptív mérésnek előfeltétele, hogy a tanulók által adott válaszokat a rendszer azonnal értékelje, pontozza, a képességfejlettséget ez alapján becsülje. Az automatikus kiértékelési rendszerben a kérdéseket egy számítógép adja a vizsgálati alanyoknak, és a kapott válaszokból összesített eredményeket szolgáltat a szakértő felé (Gergely & Takács, 2023). Ebben az értelemben az OKM (nem automatikus) kiértékelési rendszer, mely eredményként a tanuló képességpontját vagy képességszintjét szolgáltatja a visszajelző rendszerek, mérési szakértők és a pedagógusok felé. Automatizált kiértékelési rendszerek tervezésének során felmerül a nyílt végű kérdések, avagy az élőerős szakértelmet kívánó feladatok elhagyásának problémája. Az automatizált kiértékelési rendszerek esetében a tesztirányításnak nem részei az élőerős kódolást igénylő kérdések, ezeket a mérés után, utólag lehet a képességbecslés folyamatába bevonni. Fontos kérdés tehát, hogy az ilyen kérdések elhagyása mely tesztalanyok esetében és milyen következménnyel jár az értékelés eredményét, szűkebben értve a tesztirányítást tekintve.

[40] A fejezet az *Alkalmazott Pszichológia* folyóiratban megjelent cikk (T. Kárász & Takács, 2023) alapján készült.

A diákok az OKM során két tesztrészből álló, részenként nagyságrendileg 50–60 kérdésből álló tesztet töltenek ki. A kérdések jelentős része egyszerű vagy többszörös választásos feladat, kisebb része, nagyságrendileg a harmada (pl. Balkányi et al., 2018; Lak et al., 2018) nyílt végű, azaz az a válasz önálló megalkotását igényli. Nyílt végű lehet az olyan kérdés, mely nyílt kérdés (hány almát szedtünk és egy számadatot várunk), de számítógépes adatfelvétel esetén egy számítógép segítségével könnyen tudunk értékelni. Kódolandó egy nyílt végű kérdés akkor, ha mindenképpen képzett kódoló általi feldolgozásra van szükség. Ilyen lehet egy matematikai okfejtés vagy egy bizonyítás, esetleg egy fogalmazás típusú válasz a szövegértés teszten (Balázsi et al., 2014). Papír-ceruza tesztek esetében a nyílt kérdések mindkét formája kódolandó.

Az OKM esetében a nyílt, illetve zárt kérdéseknek nincsen deklarált tartalmi területük vagy gondolkodási műveletük (ld. 3.5.2 fejezet). Ez azt jelenti, hogy akár a szövegértés, akár a matematikai kompetenciák területén alkalmazhatnak nyílt kérdéseket, illetve nincsen külön olyan műveleti vagy kompetenciaterületi megosztás, mely elvárná a nyílt kérdések használatát (Balázsi et al., 2014). A tartalmi keret alapján a tesztfüzetekbe a feladatokat gondolkodási művelet és tartalmi terület/szövegtípus alapján válogatják be (Balázsi et al., 2014). Ugyanakkor a feladat formája szerint a hosszabb választ igénylő kódolandó nyílt kérdések feltételezhetően a nehezebb feladatok közé tartoznak – így valódi információs hozzájárulásuk a magasabb teljesítményű régiókban jelenik meg.

Az OKM esetében a képességpont egy folytonos, látens képességfejlettség becslése. Ugyanakkor a képességpontokat megfeleltetjük képességszinteknek is (ld. 3.5.2 fejezet). Ez a megközelítés megfelel a többszakaszos adaptív tesztelés céljának, ahol a tesztirányítás során a következő szakasz szintjét kell elsődlegesen meghatározni. A képességszintek felőli megközelítés értelemezhető a számítógépes adaptív tesztelés keretében is, ahol a mérés célja lehet szintekre történő megbízható besorolás (ld. 2.3.1 fejezet), vagy egy nagyon leegyszerűsített megvalósítás során, ahol az item nehézsége helyett az item szintje szerepel az itemkiválasztási folyamatban.

6.2.1. Minta és módszertan

Az eredmények a 2017-es főmérés tanulói szintű adatain alapulnak. A mérésben 6. évfolyamon 91 599, 8. évfolyamon 87 990, 10. évfolyamon 84 957 mérésre kötelezett diák vett részt, melyekből a hiányzások és teljes mentességek után 6. évfolyamon 85 563, 8. évfolyamon 80 833, 10. évfolyamon 76 504 diák rendelkezett kitöltött tesztfüzettel és

értékelhető eredménnyel. Közülük nem minden diák volt figyelembe vehető (pl. a sajátos nevelési igényű tanulók egy része nem mentesül a részvétel alól, de eredményük nem számít bele az aggregált eredményekbe), így végső soron a teljes elemzésbe, a mentességgel rendelkezők kizárása után 6. évfolyamon 81 647, 8. évfolyamon 77 105, 10. évfolyamon 73 728 diák adatai maradtak meg.

Az item szintű adatok segítségével diákonként kétfajta pontszámot számítottam: egyik oldalról a teljes (nyílt és zárt végű itemeket egyaránt tartalmazó feladatsorból), másik oldalról a kizárólag zárt itemeket tartalmazó, rövidebb tesztből álló kérdéssorból számított teljesítmény pontszámokat használva. Ekkor minden diák esetében rendelkezem egy olyan pontszámmal, amely esetében a nyílt válaszai is kódolásra kerülnek, valamint egy olyannal, amelynek számítása során automatikus kiértékelést kértem a pontszámító rendszertől. Nyílt végű itemnek nevezünk minden olyan itemet, amely szabad szöveges választ kíván és a tesztfüzetek feldolgozása során képzett kódolók értékelik az eredményt. Ebben az értelemben nyílt végű itemnek számítottak azok a feladatok is, amelyekben egyetlen szó vagy szám a válasz, habár ezeket a számítógépes tesztekben lehetséges automatikus itemként kezelni (pl. példaválaszok gyűjteményével vagy bizonyos adatbeviteli feltételek rögzítésével).

Az Országos kompetenciamérés sajátossága, hogy a felmért évfolyamokon teljes körű, lényegében az aktuális populációt méri (Belinszki Bálint et al., 2020). A képességpontokra az Országos kompetenciamérés módszertanának megfelelő módon súlyozást alkalmaztam (Auxné Bánfi et al., 2014), azaz az országos eredménybe beleszámító, de hiányzó tanulók eredményét az osztályátlaggal helyettesítettük, így a tanulói eredmények összesen 6. évfolyamon 86151, 8. évfolyamon 80833, 10. évfolyamon 76550 olyan jelentésre jogosult tanulót reprezentálnak, akiknek mindkét területen volt képességpontja. A zárt végű itemekből számított képességpontokat PARSCALE 4.1 (DuToit, 2003; Muraki & Bock, 1991) programcsomag segítségével számítottam, a további számításokat pedig IBM SPSS 28.0 programcsomagban végeztem.

A próbákat 95%-os szignifikancia szint mellett végeztem el. A képességpontok kapcsolatvizsgálatát Pearson-féle korrelációval vizsgáltam. A kereszttáblás elemzéseknél a khi-négyzet próba szignifikanciáját és a korrigált standardizált reziduálisokat is figyelembe vettem, mint kategóriánkénti hatásmértéket.

6.2.2. A teljes és csak zárt itemek alapján számított képességbecslések kapcsolata

Első lépésben megvizsgáltam a Pearson-féle korrelációt a teljes, valamint a csak a zárt végű kérdésekből számított képességpontok között (7. táblázat). A korrelációs együtthatókon megfigyelhetjük, hogy a szövegértés és a matematika képességpontok egymással 0,664 – 0,777 közötti szinten korrelálnak, függetlenül attól, hogy az adott területet teljes teszttel vagy csak zárt itemekkel mértem. Ez nagyságrendben azonos, közepes vagy erős kapcsolatot mutat (Vargha, 2015) a két terület között, függetlenül a nyílt itemek használatától. A zárt kérdésekből számított képességpontok és a teljes tesztből számított képességpontok mindkét területen 0,9 feletti, de 1-nél kisebb korrelációt mutatnak, ami nagyon erősnek számít. Ezt értelmezhetjük úgy, hogy a két teszt által mért látens képességfejlettség azonos, azaz a zárt itemek segítségével ugyanazt a képességet vizsgáljuk, amit a nyílt itemeket tartalmazó teljes teszttel. Ezt erősíti meg az is, hogy a mérési területek közötti kapcsolat hasonló nagyságú, tekintet nélkül a feladatformára.

7. táblázat

A teljes teszt és a csak zárt végű itemek alapján számított képességpontok Pearson korrelációs együtthatói

Pearson korreláció		1.	2.	3.	4.
1. Matematika képességpont, teljes teszt	6. évf.				
	8. évf.	–			
	10. évf.				
2. Szövegértés képességpont, teljes teszt	6. évf.	0,723**			
	8. évf.	0,777**	–		
	10. évf.	0,775**			
3. Matematika képességpont, zárt itemek	6. évf.	0,910**	0,674**		
	8. évf.	0,954**	0,739**	–	
	10. évf.	0,963**	0,741**		
4. Szövegértés képességpont, zárt itemek	6. évf.	0,703**	0,932**	0,664**	
	8. évf.	0,741**	0,958**	0,716**	–
	10. évf.	0,752**	0,951**	0,729**	

Megjegyzés. $N_6 = 86151$, $N_8 = 80833$, $N_{10} = 76550$.
**: $p < 0,01$.

6.2.3. A teljes és csak zárt itemek alapján becsült képességszint összehasonlítása

A teljesítménypontok együttjárása után a matematika és szövegértés területeken a kétféle pontszámításból adódó szinteket hasonlítottam össze, hogy kiderüljön, a nagyobb képet vizsgálva milyen szinteken milyen irányokban torzítanak a pontszámok variánsai. Ezt a fajta egyéni eltérést árnyalja az, ha a diákoknak nem a pontszámát, hanem a szintjét igyekszünk megragadni. Az Országos kompetenciamérés képességskálája 1500 pontos átlaggal és 200 pontos szórással került kialakításra, ami 1200 és 1800 pontszámok közötti értékeket valószínűsít. A képességskálát 8 szintre osztják, egy szint megközelítőleg 100 pont terjedelmű. Emellett a diákok teljesítményének standard hibája (az egyéni mérés hibájának, mint valószínűségi változónak a szórása) nagyságrendileg 50–80 pont körül van, tehát akkor számíthatunk a képességszint téves azonosítására, ha a tanuló pontszámát tekintve két szint határán mozog.

Matematikából mindhárom évfolyamon megfigyelhető, hogy a magasabb képességszinteken a nyílt itemek elhagyásával mindkét irányú torzítás megfigyelhető (8**. táblázat**-10. táblázat)

A szint változásának aránya 20% körül van mind a magasabb, mind az alacsonyabb szintre sorolás esetén. A magasabb szinteken a torzító hatás a jobb teljesítményt bünteti abban az értelemben, hogy csak a zárt kérdésekkel dolgozva a diákok teljesítménye a teljes teszt eredményéhez képest romlik, azaz a jobb teljesítményű diákok jó teljesítménye elsősorban a jellemzően nehezebb, nyílt végű kérdéseken tud igazán kiteljesedni. Ez azt is jelenti azonban, hogy már 5. és 6. képesség-szinteken, tehát matematikai területen némileg az átlagos tudásszint felett jelentkezik a nyílt kérdések szükségessége. Ez a különbséget adó eltérés azt jelenti, hogy ha adott két diák, akiknek vizsgáljuk a matematika teljesítményét teljes pontszám és zártakkal való számítás esetében, akkor, ha *A* diák jobban teljesít a teljes mérés tekintetében *B* diáknál, akkor a nyílt kérdések elhagyása mellett nem tudja vezető szerepét megőrizni.

Az alacsonyabb képességszintet elérő tanulók jellemzően jobban teljesítenének a nyílt kérdések elhagyásával. Az első alatti szinten teljesítők 40%-a és az 1. szinten teljesítők harmada a következő képességszintre kerülne.

8. táblázat

A teljes teszt és a csak zárt itemek alapján számított képességgponthoz tartozó képességszint összehasonlítása 6. évfolyamon matematikai eszköztudás területen

Matematika képesség-szint, teljes teszt alapján		Matematika képesség-szint, zárt teszt alapján								Összes
		1. alatt	1.	2.	3.	4.	5.	6.	7.	
1. alatt	N	**1936**	*1554*	2	0	0	0	0	0	3492
	AR	**191,2**	*64,6*	-33,0	-37,1	-30,5	-20,8	-11,6	-6,0	
1.	N	*486*	**6051**	*2994*	28	0	0	0	0	9559
	AR	*14,0*	**173,3**	*20,1*	-63,1	-52,5	-35,8	-19,9	-10,3	
2.	N	21	1829	**13557**	*4118*	46	0	0	0	19571
	AR	-26,2	-8,4	**174,1**	*-22,9*	-79,6	-54,9	-30,5	-15,9	
3.	N	0	43	3326	**16641**	4505	132	0	0	24647
	AR	-31,7	-64,3	-42,5	**166,8**	-9,8	-60,9	-35,6	-18,5	
4.	N	0	0	54	2852	**11521**	*3408*	180	24	18039
	AR	-25,8	-53,1	-81,8	-39,4	**162,9**	*40,4*	-20,8	-13,0	
5.	N	0	0	0	17	1504	**5165**	*1391*	147	8224
	AR	-16,3	-33,5	-52,3	-58,2	-5,0	**161,3**	*69,0*	7,8	
6.	N	0	0	0	0	4	*462*	**1371**	*368*	2205
	AR	-8,1	-16,7	-26,1	-29,3	-23,9	*15,9*	**151,0**	*75,7*	
7.	N	0	0	0	0	0	0	*106*	**308**	414
	AR	-3,5	-7,2	-11,2	-12,5	-10,3	-7,0	*24,4*	**151,8**	
Összes	N	2443	9477	19933	23656	17580	9167	3048	847	86151

Megjegyzés. AR azt jelzi, hogy a megfigyelt gyakoriság alacsonyabb (negatív AR) vagy magasabb (pozitív AR), mint az elvárt gyakoriság. A 2-nél nagyobb vagy -2-nél kisebb értékek már eltérést jeleznek.

N = gyakoriság; AR = korrigált standardizált reziduális.

Megjegyzés. A helyes kategorizálást vastagítás jelezi. Dőlt betűvel jelezve, ha a zárt itemekből álló teszt alapján az egy szintnyi tévedés az AR alapján pozitív és jelentősnek mondható, azaz a vártnál lényegesen több eredmény jelenik meg.

9. táblázat

A teljes teszt és a csak zárt itemek alapján számított képességgponthoz tartozó képességszint összehasonlítása 8. évfolyamon matematikai eszköztudás területen

Matematika képességszint, teljes teszt alapján		Matematika képességszint, zárt itemek alapján								Összes
		1. alatt	1.	2.	3.	4.	5.	6.	7.	
1. alatt	N	**828**	*704*	0	0	0	0	0	0	1532
	AR	**180,1**	*65,4*	-16,0	-21,1	-22,7	-19,0	-12,7	-7,2	
1.	N	*250*	**2977**	*1387*	8	0	0	0	0	4622
	AR	*24,6*	**169,7**	*32,0*	-37,1	-40,2	-33,7	-22,4	-12,8	
2.	N	17	*1243*	**7274**	2115	28	0	0	0	10677
	AR	-11,5	*25,3*	**172,1**	-6,4	-63,0	-53,4	-35,5	-20,3	
3.	N	0	51	2685	**11893**	3271	74	3	0	17977
	AR	-17,8	-37,1	3,6	**160,9**	-23,3	-71,6	-48,6	-27,8	
4.	N	0	0	62	3881	**13509**	3366	109	6	20933
	AR	-19,7	-43,0	-66,7	-14,8	**154,6**	-11,8	-50,8	-30,5	
5.	N	0	0	0	43	3240	**10124**	*2280*	107	15794
	AR	-16,4	-35,9	-56,8	-73,9	-14,0	**162,3**	*24,7*	-20,3	
6.	N	0	0	0	0	11	*1654*	**4613**	*883*	7161
	AR	-10,4	-22,7	-35,9	-47,3	-50,6	*9,7*	**168,2**	*45,6*	
7.	N	0	0	0	0	0	0	*517*	**1620**	2137
	AR	-5,5	-12,0	-19,0	-25,0	-26,9	-22,6	*24,0*	**192,1**	
Összes	N	1095	4975	11408	17940	20059	15218	7522	2616	80833

N = gyakoriság, AR = korrigált standardizált reziduális.

Megjegyzés. A helyes kategorizálást vastagítás jelezi. Dőlt betűvel jelezve, ha a zárt itemekből álló teszt alapján az egy szintnyi tévedés az AR alapján pozitív és jelentősnek mondható, azaz a vártnál lényegesen több eredmény jelenik meg.

10. táblázat

A teljes teszt és a csak zárt itemek alapján számított képességgponthoz tartozó képességszint összehasonlítása 10. évfolyamon matematikai eszköztudás területen

Matematika képességszint, teljes teszt alapján		Matematika képességszint, zárt itemek alapján								Összes
		1. alatt	1.	2.	3.	4.	5.	6.	7.	
1. alatt	N	**623**	*535*	0	0	0	0	0	0	1158
	AR	**174,5**	*66,0*	-12,4	-17,0	-19,9	-17,4	-12,3	-8,4	
1.	N	*188*	**2239**	*1270*	9	0	0	0	0	3706
	AR	*24,0*	**161,3**	*44,4*	-30,6	-36,2	-31,7	-22,4	-15,3	
2.	N	19	*890*	**5283**	1846	16	0	0	0	8054
	AR	-7,8	*27,4*	**160,5**	7,4	-54,5	-48,1	-34,1	-23,2	
3.	N	0	48	2188	**9608**	2586	40	2	0	14472
	AR	-14,0	-28,1	15,0	**156,2**	-22,4	-66,9	-48,0	-32,7	
4.	N	0	2	85	3641	**13153**	2496	38	2	19417
	AR	-16,9	-36,4	-56,0	-4,2	**158,4**	-30,4	-57,0	-39,4	
5.	N	0	0	0	38	3472	**11141**	1717	21	16389
	AR	-15,1	-32,6	-52,2	-70,9	-13,2	**170,0**	-4,4	-34,6	
6.	N	0	0	0	0	7	*1984*	**6286**	*1031*	9308
	AR	-10,8	-23,3	-37,2	-51,1	-59,5	*2,2*	**181,2**	*24,1*	
7.	N	0	0	0	0	0	0	*722*	**3278**	4000
	AR	-6,8	-14,7	-23,5	-32,3	-37,6	-33,0	*13,4*	**214,4**	
Összes	N	830	3714	8826	15142	19234	15661	8765	4332	76504

N = gyakoriság, AR = korrigált standardizált reziduális.

Megjegyzés. A helyes kategorizálást vastagítás jelezi. Dőlt betűvel jelezve, ha a zárt itemekből álló teszt alapján az egy szintnyi tévedés az AR alapján pozitív és jelentősnek mondható, azaz a vártnál lényegesen több eredmény jelenik meg.

Szövegértés esetében kisebb a szerepe a nyílt kérdéseknek, de hasonló jelenséget figyelhetünk meg, ugyanis ezen a területen a torzító hatás elsősorban a jobb teljesítmény tetején (6. és 7. szint) jelentkezik (11**. táblázat**–Megjegyzés. A helyes kategorizálást vastagítás jelezi. Dőlt betűvel jelezve, ha a zárt itemekből álló teszt alapján az egy szintnyi tévedés az AR alapján pozitív és jelentősnek mondható, azaz a vártnál lényegesen több eredmény jelenik meg.

13. táblázatok). A képességskála alacsonyabb régiójában nagyobb a felfelé (jobb teljesítmény felé) torzítás, míg a magasabb képességszinteken jellemzőbb a lefelé torzítás a matematika területhez képest. Azt is mondhatjuk, hogy a szövegértés esetében a torzító hatás később jelenik meg – ha úgy tetszik, akkor nagyobb képesség-szint skálát tudunk zárt itemekkel elfogadható szinten mérni.

A kereszttáblák esetében azt várjuk, hogy a főátlóban (vastagítással jeleztük), a bal felső sarkot a jobb alsó sarokkal összekötő cellákban legyenek nagy értékek, majd innen távolodva egyre kevesebb esetet találjunk. Továbbá bármelyik szinten a zárt itemek esetén a teljes teszt eredményétől 2 szintet eltérni csak igen ritkán (1% alatt) térnek el az eredmények. Egy szintnyi tévedést mindkét területen láthattuk: a matematika esetében felfelé torzítást az alsóbb régióban (lefelé torzítást a felsőben), a szövegértés esetében pedig a nagyobb torzítások inkább a felsőbb szinteken valósultak meg.

11. táblázat

A teljes teszt és a csak zárt itemek alapján számított képességponthoz tartozó képességszint összehasonlítása 6. évfolyamon szövegértés területen

Szövegértés képességszint, teljes teszt alapján		Szövegértés képességszint, zárt itemek alapján								Összes
		1. alatt	1.	2.	3.	4.	5.	6.	7.	
1. alatt	N	**760**	*503*	0	0	0	0	0	0	1263
	AR	**202,5**	*50,0*	-15,0	-19,5	-21,2	-17,2	-10,7	-5,1	
1.	N	*188*	**4017**	*1413*	0	0	0	0	0	5618
	AR	*16,7*	**210,3**	*22,0*	-42,2	-45,9	-37,2	-23,2	-11,0	
2.	N	2	804	**9676**	2482	10	0	0	0	12974
	AR	-12,9	0,1	**206,1**	-11,0	-73,0	-59,3	-36,9	-17,6	
3.	N	0	0	1840	**14716**	3615	101	5	0	20277
	AR	-17,2	-41,8	-27,1	**192,8**	-30,2	-76,0	-48,5	-23,2	
4.	N	0	0	0	2499	**16216**	3464	123	8	22310
	AR	-18,3	-44,5	-72,9	-48,2	**185,0**	-14,1	-48,4	-24,3	
5.	N	0	0	0	0	2520	**11354**	*2061*	93	16028
	AR	-14,8	-36,0	-59,0	-76,4	-32,8	**187,6**	*23,8*	-14,1	
6.	N	0	0	0	0	0	1197	**4454**	*750*	6401
	AR	-8,8	-21,3	-34,9	-45,3	-49,2	0,0	**186,1**	*58,0*	
7.	N	0	0	0	0	0	0	*420*	**860**	1280
	AR	-3,8	-9,3	-15,1	-19,6	-21,3	-17,3	*32,3*	**168,5**	
Összes	N	950	5324	12929	19697	22361	16116	7063	1711	86151

N = gyakoriság, AR = korrigált standardizált reziduális.

Megjegyzés. A helyes kategorizálást vastagítás jelezi. Dőlt betűvel jelezve, ha a zárt itemekből álló teszt alapján az egy szintnyi tévedés az AR alapján pozitív és jelentősnek mondható, azaz a vártnál lényegesen több eredmény jelenik meg.

12. táblázat

A teljes teszt és a csak zárt itemek alapján számított képességponthoz tartozó képességszint összehasonlítása 8. évfolyamon szövegértés területen

Szövegértés képességszint, teljes teszt alapján		Szövegértés képességszint, zárt itemek alapján								Összes
		1. alatt	1.	2.	3.	4.	5.	6.	7.	
1. alatt	N	**286**	*244*	0	0	0	0	0	0	530
	AR	**175,4**	*49,1*	-8,5	-11,9	-13,6	-12,2	-8,4	-4,5	
1.	N	*114*	**2190**	*780*	0	0	0	0	0	3084
	AR	*25,7*	**192,1**	*23,6*	-29,1	-33,2	-29,8	-20,6	-11,1	
2.	N	3	*848*	**6557**	1354	8	0	0	0	8770
	AR	-6,5	*28,2*	**193,3**	-13,4	-58,0	-52,2	-36,1	-19,4	
3.	N	0	1	2224	**11691**	2090	56	2	1	16065
	AR	-10,0	-29,1	8,8	**180,5**	-40,9	-73,3	-51,5	-27,6	
4.	N	0	0	8	3858	**14460**	2706	97	12	21141
	AR	-12,0	-34,8	-61,8	-11,1	**165,8**	-36,5	-59,2	-32,5	
5.	N	0	0	0	13	4155	**12193**	*2338*	118	18817
	AR	-11,1	-32,2	-57,4	-80,3	-12,7	**163,8**	*3,4*	-25,4	
6.	N	0	0	0	0	11	*2581*	**6123**	*1132*	9847
	AR	-7,5	-21,8	-38,8	-54,5	-61,9	*11,6*	**166,1**	*44,0*	
7.	N	0	0	0	0	0	1	*921*	**1710**	2632
	AR	-3,7	-10,7	-19,1	-26,8	-30,6	-27,4	*37,7*	**169,9**	
Összes	N	403	3283	9569	16916	20724	17537	9481	2973	80886

N = gyakoriság, AR = korrigált standardizált reziduális.

Megjegyzés. A helyes kategorizálást vastagítás jelezi. Dőlt betűvel jelezve, ha a zárt itemekből álló teszt alapján az egy szintnyi tévedés az AR alapján pozitív és jelentősnek mondható, azaz a vártnál lényegesen több eredmény jelenik meg.

13. táblázat

A teljes teszt és a csak zárt itemek alapján számított képességponthoz tartozó képességszint összehasonlítása 10. évfolyamon szövegértés területen

Szövegértés képességszint, teljes teszt alapján		Szövegértés képességszint, zárt itemek alapján								Összes
		1. alatt	1.	2.	3.	4.	5.	6.	7.	
1. alatt	N	**331**	*220*	0	0	0	0	0	0	551
	AR	**180,7**	*47,7*	-7,2	-10,6	-13,5	-13,5	-9,9	-5,9	
1.	N	*122*	**1599**	*536*	0	0	0	0	0	2257
	AR	*29,9*	**180,3**	*26,0*	-21,8	-27,7	-27,6	-20,3	-12,0	
2.	N	10	*742*	**3898**	929	31	0	0	0	5610
	AR	-4,3	*42,5*	**168,7**	-0,8	-43,7	-44,6	-32,8	-19,4	
3.	N	0	18	2132	**8067**	1945	76	6	1	12245
	AR	-9,4	-21,6	37,8	**157,4**	-24,9	-67,4	-50,7	-30,1	
4.	N	0	0	34	3942	**12206**	2711	82	7	18982
	AR	-12,4	-29,7	-47,8	16,1	**145,3**	-38,4	-65,0	-39,4	
5.	N	0	0	0	44	4763	**12433**	2313	91	19644
	AR	-12,7	-30,4	-49,9	-72,5	-2,1	**145,4**	-14,9	-37,4	
6.	N	0	0	0	0	43	*3680*	**7338**	*1185*	12246
	AR	-9,4	-22,5	-37,1	-54,6	-68,4	*14,9*	**151,5**	*19,6*	
7.	N	0	0	0	0	0	20	*1787*	**3208**	5015
	AR	-5,7	-13,7	-22,5	-33,1	-42,1	-41,3	*42,1*	**181,1**	
Összes	N	463	2579	6600	12982	18988	18920	11526	4492	76550

N = gyakoriság, AR = korrigált standardizált reziduális.

Megjegyzés. A helyes kategorizálást vastagítás jelezi. Dőlt betűvel jelezve, ha a zárt itemekből álló teszt alapján az egy szintnyi tévedés az AR alapján pozitív és jelentősnek mondható, azaz a vártnál lényegesen több eredmény jelenik meg.

6.2.4. Összegzés

Az elemzés elsősorban azt a kérdést járta körbe, hogy egy automatikus kiértékelő rendszer nagyobb méretű alkalmazás esetében, a nyílt végű kérdések elhagyásával (ha úgy tetszik, az élőerős kiértékelés átcsoportosításával) milyen torzításokra számíthatunk (Brassil & Couch, 2019; Bridgeman, 1992). Vizsgálatom során azt teszteltem, hogy nyílt és zárt kódolású kérdések együttes alkalmazása helyett kizárólag zárt itemek alkalmazásával azonos döntések születnek-e az egyes kategóriákba történő soroláskor.

Az OKM esetében a nyílt itemek nem egyes gondolkodási műveletek vagy tartalmi területek pontosabb mérése miatt szerepelnek, hanem a mérés egészének

változatosságához járulnak hozzá (Balázsi et al., 2014). Ugyanakkor a pedagógiai munka vagy a munkahelyi kiválasztási szituációban továbbra is a folyamat elengedhetetlen része a nyílt végű kérdés és az interjú. Ennek megfelelően kutatási kérdés arra irányult, hogy az értékelési folyamat mely pontján vagy mely személyeknél szükséges a nyílt kérdések használata.

A számítások igazolták, hogy a folytonos kiértékelések során meglehetősen közeli, 0,9 feletti korrelációs szintű egyezést tudtunk kimutatni a teljes teszt és a csak zárt itemek alapján számított képességpontok között. Fontos azon feltétel teljesítése, hogy az Országos kompetenciamérés esetében meglehetősen jó minőségű és többszörösen tesztelt kérdésekkel dolgozzanak a szakemberek (Auxné Bánfi et al., 2014), ami garanciát jelent arra is, hogy néhány kérdés elhagyása nem okoz rendszer szintű problémákat.

A folytonos teljesítményt az OKM módszertanának megfelelően képességszintekbe osztva, majd a két teljesítményből számított képességszintet összevetve kiderül, hogy a csak zárt kérdések segítségével készült szintekre való besorolás 8 képességszint esetén 2 szintnél nagyobb arányú tévesztést az esetek kevesebb, mint 1 százalékában eredményez. A kategória szintű vizsgálatok azt mutatták, hogy jellemzően a mérési skálánk két végletén tapasztalhatók jelentősebb eltérések, ami egybevág Geer (1991) eredményeivel. A magasabb képességszinteken teljesítők esetében lefelé, a legalsó képességszinteken teljesítők esetében felfelé torzítás észlelhető a nyílt végű itemek (élőerős kiértékelést igénylő feladatok) elhagyásával. Ez alapján a szöveges választ kívánó feladatok elhagyása a felsőbb tartományban a valóban jó képességgel rendelkezőket némileg alulértékeli, ugyanakkor a teljesítmény skálák alsó régióiban éppen ezzel ellentétes torzításokkal jártak együtt, a rosszabb képességű kitöltők hiányosságai rejtve maradhatnak.

Ennek fényében az állapítható meg, hogy a szakemberek szerepe továbbra sem hagyható el az értékelési vagy kiválasztási folyamatok során, ahogy az osztálytermi értékelésben sem elhagyható a pedagógusok szakértelme. Az eredmények azt mutatják, hogy a szakemberek bevonása adott esetben a folyamatok későbbi pontjaira tehető abban az értelemben, hogy a zárt itemekből alkotott tesztek a nagyon alacsony vagy nagyon magas teljesítményt jól kimutatják, így osztálytermi környezetben nagyobb valószínűséggel lehet olyan diákokkal időigényesebb feladatokat végezni, akiknél nagyobb hiányosságok vagy tehetségek lehetnek fellelhetők.

A zárt itemek használata előfeltétele az azonnali eredmény számításának, egyszersmind az adaptív mérésnek. A nyílt itemek egy része ugyanakkor a papír-ceruza

teszt formához hasonló zárt itemmé alakítható, ami például a 2022. évi OKM esetében meg is történt. A korábban nyílt végű sorbarendezés feladattípus jól átalakítható automatikusan értékelhető „fogd és vidd” feladattá, az egyetlen szám beírását igénylő nyílt végű feladatok pedig automatikus kódolásúak (Balázsi et al., 2021).

6.3. Elméleti optimum[41]

Az előző két alfejezetben ismertetett eredmények megalapozták a papír-ceruza teszt adatai és a szimulációk közötti kapcsolatot. A jelen alfejezetben ismertetett eredmény– és annak speciális esete – a megbízhatóság fogalmára alapul. Az adaptív mérési rendszereket egyedi teszteknek tekintem, melyek különböző célokkal rendelkeznek, ezért a megbízhatóságot és a rendszerek tulajdonságait a 2.3 fejezetben meghatározott elemekből tervezzük. Ugyanakkor a levezetés és az abból származó képletek és illusztrációk megteremtik a hidat az elméleti modellek egyenletei és a gyakorlat között.

A levezetéshez rögzíteni kell néhány feltételt. Általában véve abból a feltevésből indulok ki, hogy létezik egy tulajdonságait tekintve jól ismert lineáris teszt (ezen a ponton mindegy, hogy papír-ceruza vagy számítógépes formában), és ezt a tesztet szeretnénk számítógépes adaptív megvalósításra átültetni. Ekkor a feltételek legyenek a következők.

1) A teszt pontosságát adottnak tekintjük, azaz itemek és tanulók szintjén ugyanazt a becslési pontosságot szeretnénk elérni, ami a lineáris teszttel elérhető.
2) A teszt kizárólag két értékű itemekből áll (helyes/helytelen válasz). Ez súlyos megszorításnak tűnhet, de a korábbi fejezetekben bemutatott tanulóiteljesítmény-mérések legnagyobbrészt ilyen itemeket alkalmaznak.
3) A lineáris teszt itemeinek paramétereit ismerjük, olyan itemekkel rendelkezünk, melyeknek elsősorban a nehézsége ismert a tesztet összeállítók előtt. Ez nem túlzó feltételezés, mert a szakemberek ezzel az információval rendelkezni szoktak a tesztek összeállításakor (valamilyen próbamérés után), a nagymintás főmérés után még pontosabban. Erre példa az eltérő itemműködés vizsgálata (pl. OECD, 2017b). Harrison és munkatársai (2017) munkájában még az is tisztázott, hogy egy-egy feladat gyermekek és felnőttek számára mennyire nehéz – tehát a szakértők akár korosztályonként is meglehetősen jó becsléseket tudnak mondani a feladatok nehézségét illetően.

[41] A fejezet az *Alkalmazott Matematikai Lapokban* megjelent cikk (T. Kárász & Takács, 2021) alapján készült.

Kutatási kérdésem: ugyanazon jelenséget vizsgálva, az eredeti lineáris teszttel egyező típusú kérdésekkel ugyanolyan pontosságot mennyivel gyorsabban (kevesebb itemmel) tudunk elérni? A becslési pontosságon egy-egy egyénre vonatkoztatott becslési pontosságot, a becslési hiba nagyságát értem. Az adaptív eljárás alkalmazásával nem a teljes mintafelvétel hibáját szeretném csökkenteni (ami egy lehetséges cél az eredeti teszthossz megtartásával), hanem az adott teszten elérhető pontszám hibáját szeretném elérni – kevesebb item felhasználásával.

A kutatókat általánosságban érdeklő és foglalkoztató kérdés az adaptív tesztelés során inkább az szokott lenni, hogy mekkora itembankra van szükség ahhoz, hogy egy adaptív rendszert működtetni lehessen (Magyar, 2014b). Ha az IQ-t szeretnénk 100 kérdés helyett csak 10 kérdésből mérni, amikor azok nyilvánosságra kerülnek, korrumpálódnak (pl. Cizek & Wollack, 2016), akkor nagyon gyorsan használhatatlanná válik a teszt. Ezt azzal lehet kivédeni, hogy nagy méretű itembankot hozunk létre, amiből a megkérdezett véletlenszerűen kap kérdéseket – az aktuális szintjének megfelelően.

Bár a megbízhatóság becslési formulái meglehetősen régóta ismertek (Cronbach, 1951; Kuder & Richardson, 1937; Wright, 1977) (ld. 2.1 fejezet), az adaptív tesztelés esetében a megközelítés általában inkább szimulációs módszerek alkalmazását jelentette (ld. 5.1 fejezetben). A szimulációk során olyan „mi lenne ha” szcenáriókat vizsgálnak, hogy adott feltételek (pl. mintanagyság, képességeloszlás, itembank, teszthossz) esetében hány kérdésre lenne szükség az adott becslési szint teljesítéséhez.

6.3.1. Kuder-Richardson formula

A tesztek során alapvető definíciónak a teszt reliabilitását veszem. Az elméleti reliabilitás becslése a Kuder-Richardson formula szerint az alábbi alakban határozható meg (Kuder & Richardson, 1937, (20) képlet):

$$KR-20 = \frac{L}{L-1}\left(1 - \frac{\sum_{i=1}^{k} p_i\, q_i}{s^2}\right) = r,$$

ahol L az itemek száma, $\sum_i p_i q_i$ az itemek varianciájának összege, p_i az összes jó válasz aránya (jó válasz / összes esetszám), míg q_i a rossz válaszok aránya, továbbá s^2 a teljesítmények varianciáinak összege, amely s^2 alap esetben $N(0,1)$ (azaz 0 várható értékű és 1 szórású, úgynevezett standard normális eloszlású) változók négyzetösszegét jelenti. Ha a mintaalanyok egymástól függetlenül írják a tesztet, akkor s^2 legalábbis nagyságrendileg a minta nagyságát jelenti.

A teszt a minta növelésével egyre megbízhatóbbá válik, hiszen a KR-20 második tagjában a hányados határértékben nullához konvergál. Feltéve, hogy valóban standard normális határeloszlást tudunk az IRT modellek segítségével meghatározni minden résztvevő pontszámaként. A nemzetközi és hazai mérések egyaránt erre a feltételezésre építenek. A formula definíciói alapján a teszt standard hibáját (*standard error of measurement,* SEM) az alábbi formula szerint definiálják:

$$SEM = s\sqrt{1-r}.$$

Ha a teszt itemei megoldhatóságukat tekintve kiegyensúlyozottak, azaz a teszt teljes varianciájához képest az itemek varianciája összességében alacsony, akkor a $(\sum pq)/s^2$ kifejezés értéke alacsony lesz. Amennyiben az itemek összvarianciája a teljes teszt/teljesítmények varianciájához képest alacsony marad, úgy az $1-(\sum pq)/s^2$ kifejezés egyre jobban közelít 1-hez. Ebből következően tehát, relatíve hosszú tesztek esetében így a teljes teszt reliabilitása 1-hez közelít. Minél közelebb van a teszt reliabilitása 1-hez, várhatóan annál kisebb lesz az $s\sqrt{1-r}$ kifejezés értéke, tehát annál kisebb lesz a teszt standard hibája.

Ezen a ponton szokás a szimulációkat végrehajtani. Az egyes itemek megoldottsága természetesen nem azonos, vannak a tesztekben könnyebb és nehezebb feladatok, tehát a teljes tesztre ránézve azt tudjuk szimulálni, hogy egyik-másik itemet elhagyva (vagy bevéve az eljárásunkba), mennyivel tudunk gyorsabban célba érni – relatíve kevesebb itemet felhasználva hasonló eredményre jutni.

A helyzet azonban az, hogy adaptív tesztelés esetében ez nem egészen így történik, ezt van der Linden és Glas (2000) az elsők között említették. Több itembank nagysággal és teszthosszal, mintanagysággal is végeztek szimulációs vizsgálatot.

Az egyik lehetséges cél tehát az, hogy rögzítve a teszt reliabilitását, a teljes minta hibájának mértékét, azt szeretnénk megtudni, hogy milyen itemszámra van szükségünk, ha adaptív módon szeretnénk tesztelni, feltéve, hogy az esetszám (mintanagyság), melyet a becsléshez felhasználunk, változatlan marad.

Tekintettel arra, hogy a SEM formulája nem tartalmazza azokat a (képesség)szinteket, melyekkel a nemzetközi mérések (pl. OECD, 2019a), illetve az OKM (Balázsi et al., 2014) jól interpretálható értelmezést adnak a képességpontok jelentésének, így olyan formulával dolgoztam helyette, mely ezt a sajátosságot figyelembe veszi. A szintek esetében az alábbi más példákra is gondolhatunk:

1) Betegség esetében egy adott betegség súlyosságának fokozatai.

2) Iskolai teljesítmény esetében nem egy teszt pontszámait értjük alatta, hanem az arra kapott osztályzatot.

6.3.2. Wright formulája – általános eset (0 < p < 1)

Számítógépes adaptív tesztelés (CAT) esetében a p = 1/2 esetet (ahol p a jó megoldás valószínűsége) érdemes elemezni, mivel az adaptív mérés folyamata arra épül, hogy lehetőleg olyan itemet kap a teszt kitöltője, hogy ugyanolyan valószínűséggel (50%) oldja vagy nem oldja meg a feladatot. Ettől eltérve, jelen elemzés abból indul ki, hogy általánosságban könnyebb ($p > 0{,}5$), illetve nehezebb ($p < 0{,}5$) tesztek is górcső alá vehetők. Ez az általánosítás jogos lehet, mivel az adaptív tesztek motivációval való kapcsolatát vizsgáló kutatások eredményei vegyesek, a könnyebb itemekből álló tesztek azonban motiválóbbak lehetnek (Akhtar et al., 2023). Fontos azonban kiemelni, hogy a formula p és q esetére szimmetrikus, és most a $p \leq 0{,}5$ eseteket fogjuk formalizálni. Világos, hogy a $p = 0$ (a feladatot nem lehet megoldani) és a $p = 1$ (a feladatot mindenki meg tudja oldani) esetek nem érdekesek.

Az előzőekben megismert jelölések az alábbi alakban írhatók át. Tegyük fel, hogy b_i jelöli az adott személy képességfejlettségét ($i = 1, \dots, N$ kitöltővel számolva) és d_j jelöli az adott itemek nehézségét ($j = 1, \dots, L$ itemet használunk a papír-ceruza referencia tesztben). Jelölje s_j azt a számot, ahányan az adott itemet jól megoldották (valamint n_w jelölje azok számát, akik pontosan w darab feladatot oldottak meg helyesen). Ha adaptív tesztelésben gondolkodunk, akkor a korábbi jelölésekkel $s_j = Np$ (illetve $s_j = \frac{N}{2}$ a hagyományos CAT esetében). Továbbá nem életszerűtlen az a megközelítés, hogy azokat az alanyokat, akik minden itemet megoldanak vagy elrontanak, kihagyjuk a további elemzésekből (a kitöltők számát továbbra is N-nel jelölve). Ezeknek a kitöltőknek a képességbecslése pl. maximum likelihood eljárásokkal nem is lehetséges, ezért az IRT alapokon működő tesztek esetében valamilyen előre meghatározott korlátozó értékkel helyettesítik a képességpont becslését. Hasonlóan, ha egy itemet mindenki megoldott/senki sem oldott meg, szintén elhagyhatjuk (jelölje a továbbiakban is a teszt hosszát L). Ez megfelel egyrészt a klasszikus tesztelmélet egyik első ellenőrző lépésének (ezek az itemek nem különböztetik meg a tesztalanyokat), másrészt a modern tesztelméletben a paramétereik nem becsülhetők meg a válaszok alapján. N és L a teszthossz és kitöltők valid számát fogja jelölni.

Wright (1977) a fenti jelölések mellett az alábbi képleteket definiálja:

$$x_j = \ln\left(\frac{N - s_j}{s_j}\right),$$

$$x = \sum_{j=1}^{L} \frac{x_j}{L},$$

$$U = \sum_{j=1}^{L} \frac{\left(x_j - x\right)^2}{L - 1}.$$

A képletek első szettje az itemekre vonatkoznak. Az első formula egy item relatív megoldottságát jelzi (x_j a j. item relatív nehézsége), amit a nem-megoldók és a megoldók aránya alapján számít. A második képlet a teljes teszt átlagos relatív nehézségét (x), a harmadik a teszt varianciáját (U) adja meg.

$$y_w = ln\left(\frac{w}{L - w}\right),$$

$$y = \sum_{w=1}^{L-1} \frac{n_w y_w}{N},$$

$$V = \sum_{w=1}^{L-1} \frac{n_w (y_w - y)^2}{N - 1}.$$

A képletek második csoportja a teszt kitöltőire vonatkozó paralel állítások. Az első képlet a pontosan *w* pont elérésének relatív nehézsége egy adott tesztalany esetében, *y* a tesztalanyok várható eredménye, *V* pedig a tesztalanyok teljesítményének varianciája. Továbbra is Wright jelöléseit használva:

$$X = \sqrt{\frac{1 + \frac{U}{2{,}89}}{1 - \frac{UV}{8{,}35}}},$$

$$Y = \sqrt{\frac{1 + \frac{V}{2{,}89}}{1 - \frac{UV}{8{,}35}}}.$$

Az X és Y értékek a teszt és a vizsgálati alanyok teljesítményét mutatják, ahonnan

$$d_j = Y\left(x_j - x\right),$$

$$SE\left(d_j\right) = y\sqrt{\frac{N}{s_j\left(N - s_j\right)}}.$$

A d_j paraméter az összes megoldóra megoldásra/teljesítményre vetítve az itemek x_j nehézsége, *SE(dj)* az adott itemek nehézségének standard hibája.

$$b_w = Xy_w$$

$$SE(b_w) = X\sqrt{\frac{L}{w(L-w)}}$$

Ezzel szemben b_w az adott megoldók, adott tesztet kitöltők átlagos teljesítménye lesz, illetve *SE(b_w)* a teljesítményeken lévő átlagos (standard) hibaként kerül bevezetésre.

Látható tehát, hogy e mutatók segítségével megadható, hogy az itemeknek, illetve a teljes tesztnek mi lesz a hibája, milyen biztonsággal tudunk item-paramétert vagy teljesítményt meghatározni.

Ha adaptív módon, de kicsit nehezítve vagy könnyítve szeretnénk mérni (tehát mindenki a saját képességfejlettségének megfelelő itemet kap, de rögzített p, illetve q valószínűséggel oldja meg/rontja el a feladatokat), valamint úgy kezeljük, hogy K szinten akarjuk az eredményeket kezelni, ahogy Wright (1977) 11 szinttel számolt, tehát -5 és 5 közötti képességértékekkel dolgozott, akkor további, szintén nem túlságosan életszerűtlen egyszerűsítések tehetők. Tudjuk, hogy

$$s_i = \frac{Np}{K},$$

ahol s_i az i itemet sikeresen megoldók száma, hiszen ilyen esetben egy adott szintet mérő itemet az arra a szintre tartozó kitöltők látnak (tehát azok adott hányada fogja megoldani).

Ez utóbbi úgy is felfogható (ezért nem életszerűtlen a megkötés), hogy egy adaptív teszt esetében a nagyon jó nem kap nagyon könnyű feladatokat és a nagyon alul teljesítő sem kap megoldhatatlannak látszó példákat. Miként egy igen súlyos állapotban lévő pácienstől sem kérdezik az enyhe tüneteket – és az alapvetően enyhébb panaszokkal érkezőket sem a rendkívül súlyos esetekre jellemző tünetek mentén kezelik. Ebből az egyszerűsítésből következik, hogy

$$x_i = \ln\left(\frac{N - \frac{Np}{K}}{\frac{Np}{K}}\right) = \ln\left(\frac{K-p}{p}\right).$$

A fenti formula szerint a jó és rossz válaszok aránya átlagosan csak a szintek számától és az adott megoldási valószínűségtől függ (minden szinten lényegében állandó, hogy hányan, milyen arányban oldják meg jól vagy rosszul a feladatokat). Ebben az esetben az is elmondható, hogy

$$x = \frac{\sum_{i=1}^{L} \ln\left(\frac{K-p}{p}\right)}{N} = \frac{L}{N}\ln\left(\frac{K-p}{p}\right),$$

ami az átlagos megoldottsági/elrontottsági kapcsolati mutató.

Ezek után a variancia:

$$U = \frac{\sum_{i=1}^{L}\left(\ln\left(\frac{K-p}{p}\right) - \frac{L}{N}\ln\left(\frac{K-p}{p}\right)\right)^2}{L-1},$$

$$y_w = ln\left(L\frac{p}{q}\right).$$

Ha adott személy mindig p valószínűséggel tudja megoldani a feladatokat, akkor a jól megoldott feladatok száma $w = Lp$, amiből következik, hogy optimális adaptív teszt esetében $V = 0$ és $Y = 1$. Optimálisan adaptív egy teszt akkor, ha valóban minden vizsgálati alany folyamatosan a számára megfelelő szinten, tehát p valószínűséggel oldható feladatokat kap. Innen viszont azt is tudhatjuk, hogy

(1) $$SE(d_i) = \sqrt{\frac{N}{s_i(N-s_i)}} = \frac{1}{\sqrt{N}}\left[\frac{K}{\sqrt{p(K-p)}}\right],$$

azaz az adott item nehézségparaméterének hibája annál nagyobb, minél több szintet szeretnénk bemérni (adott szinten kevesebb a kitöltő, így kisebb az itemről való információ), viszont minél több kitöltővel rendelkezünk, annál jobban csökken a paraméterbecslés hibája. Ez egybevág azzal az intuícióval, hogy minél szélesebb spektrumon szeretnénk, hogy egy kérdés jól mérjen, annál nagyobb bizonytalansággal tudjuk megtenni (specifikus kérdések pontosabban mérnek). Illetve, hogy a kitöltők számának növekedésével együtt jár az, hogy az itemek viselkedését egyre pontosabban fogjuk ismerni.

Szintén Wright (1977) formulái alapján megadható a teljesítmény hibája:

$$SE(b) = X\sqrt{\frac{L}{w(L-w)}} = X\sqrt{\frac{1}{Lpq}}$$

Amiből további behelyettesítéssel:

(2) $$SE(b) = \sqrt{\frac{1}{Lpq}}\sqrt{1 + \frac{\left(\frac{L}{L-1}\right)\left(\frac{N-L}{N}\right)^2 \ln^2\left(\frac{K-p}{p}\right)}{2{,}89}}$$

A fenti formulából látható, hogy adott teszthossz esetében az esetszám növelésével egy ideig csökkenthető a hiba mértéke, majd lényegében stagnálni fog, ha semmi más paraméteren nem változtatunk. Az is látható, hogy *N* növekedésével egy idő után nem fogunk tudni jobb eredmény elérni, azaz egy-egy kitöltő hibáját attól nem fogjuk tudni jobban megbecsülni, hogy rajta kívül még sokan kitöltik a tesztet.

Ez azt is jelenti, hogy ha a teszt hosszát nem növeljük, akkor az esetszám növelésével egy-egy alanyra pontosabb becsléseket nem fogunk tudni tenni. Ezt

felfoghatjuk úgy is, hogy az adott itemek egy idő után kellően pontosan bemérésre kerülnek, a belőlük nyerhető információ lényegében stagnál, tehát újabb és újabb esetek hozzávételével már nem tudunk további információkhoz jutni. Ez azt is jelenti, hogy a képességeket csak úgy tudjuk egyre pontosabban mérni, hogy a teszt hosszát növeljük, ha újabb és újabb itemeket veszünk hozzá a tesztünkhöz.

Ez alapvetően nem mond ellent annak az intuitív megfigyelésnek, melyet az egészséggel kapcsolatos diagnosztikában rögzítenek, vagy a teljesítményméréseknél tapasztalhatunk. Az egészséggel kapcsolatos teszteknél nem az történik, hogy újra és újra azonos teszteket vesznek fel (ha nem ismert a diagnózis), hanem újabb teszteket, másfajta információkat csatornáznak be. A teljesítménymérés esetében sem írja meg a diák újra és újra ugyanazt a tesztet (típusfeladatot), hanem másfajta típusokkal igyekszünk pontosabb képet kapni a tudásáról.

A kitöltők száma jellemzően minimálisan 100–200 fő, a tesztek hossza pedig ritkán megy 100 fölé, tehát az esetszám emelkedésével, a többi paramétert rögzítve, a valóságos vizsgálatokban folyamatosan javuló jóságmutatókat fogunk tapasztalni.

6.3.3. A próbamérés nagyságára és a teszthosszra vonatkozó formulák interpretációja

Az illusztráció során szintén Wright (1977) nyomvonalát követem. Esetében a kitöltők száma 50 és 500 fő között alakult itemenként. Ez realisztikus szituáció, hiszen egy-egy standardizálás során a kérdőíveket hagyományosan legalább 500 fővel szokás kitöltetni, de jellemzően ennél az online felmérések során lényegesen nagyobb minták szoktak keletkezni. Fontos azonban kiemeli, hogy Wright (1977) nem adaptív, hanem papír-ceruza tesztek esetében határozta meg ezeket a számokat – és jelen esetben éppen az a kérdés, hogy ennél kevesebb kitöltővel is el lehet-e érni hasonló eredményeket egy adaptív tesztelés során.

Jelen elemzés alapvetően két kérdésre keresi a választ:

1) Első lépésben kérdéses az, hogy egy-egy item megbízhatóságához (adott hibahatár eléréséhez) hány kitöltőre van szükség minimálisan?
2) Második lépésben kérdés az, hogy ha megvan egy megfelelő méretű és pontosságú itembank, akkor ebből az itembankból hány kérdésre van minimálisan szükség ahhoz, hogy az egyes válaszadók teljesítményét meg tudjuk határozni?

Wright nyomán (ahol a teljesítmény hibája N(0,1) eloszlású) a 0,2-es szintet határozom meg, mint elérni kívánt minimumot. Ez azt jelenti, hogy a teljes teszt esetében

az itemek megbízhatósága a teljes teszt megbízhatóságának 20% alá kell, hogy csökkenjék. A kitöltőkre vonatkozó hibát/szórást ezzel szemben 0,5-ös szinten rögzítem.

A szinteket a hagyományos OECD PISA (OECD, 2019a), illetve Országos kompetenciamérés (Auxné Bánfi et al., 2014) szintjeihez szabtam, azaz a szintek száma a szimulációkban[42] 2 és 8 között lesz, azaz $K = 2, \dots, 8$. Jellemzően e két felmérés esetében a diákok teljesítményén lévő hiba nagyságrendileg a teljesítmény szórásának 40–50%-a is lehet. Ezért maradtam a teljesítmények esetében a 0,5-ös szint elérése mellett. A második esetben, amikor a teljesítményeket a teljes teszt szintjén fogom vizsgálni, K = 3, 5, 8 esetekre mutatom be az elemzés eredményeit. Az első esetet úgy foghatjuk fel, mint az alacsony-közepes-magas (K = 3) kategóriákat. Második esetben az iskolai osztályzatokat kezelhetjük szintekként. A harmadik, K = 8 esetben pedig az OKM esetét vehetjük alapnak.

A szimulációkban N = {10,20,30,50,100,200,300,400,500} esetszámokkal fogok dolgozni. A teszt hosszát L = {10,30,60} itemre állítom be. A teszt nehézségét pedig a megoldási valószínűségekkel p = {0,1;0,2;0,3;0,4;0,5} szintekre.

Első lépésben górcső alá kerül, hogy az itemek hibája miként alakul a szintek száma, a megoldottsági valószínűségek és a kitöltők száma alapján. Az eredményekből (14. táblázat) leolvasható, hogy p = 0,1 illetve p = 0,2 (mely egyéb iránt megegyezik a p = 0,9 és a p = 0,8 esetekkel) legalább 400, inkább 500 fő kell ahhoz, hogy a hibát tekintve elérjük a 0,2-es alsó határt. Ez jellemzően a nagyon nehéz/nagyon könnyű feladatok világa, amely esetben valóban több esetre van szükség a megfelelő minőség garantálásához. Ráadásul ez a szintek számának emelkedésével még nehezebbé is válik – tehát minél több szintet kalibrálunk (minél szofisztikáltabban szeretnénk mérni), annál több kitöltőre van szükség a szélsőséges feladatok pontos bemérésére.

Ezzel szemben, ha megnézzük a 0,4-es, illetve 0,5-ös szinteket, tehát a kiegyensúlyozottabb teszteket (az itemeket nagyjából fele-fele arányban oldják vagy nem oldják meg), ilyen esetekben már jellemzően 100, illetve 200 kitöltő is elegendő a megfelelő szint biztosítására (ez alól csak a 8 szint esetében van kivétel).

[42] A szimuláció ebben a fejezetben nem az 5.1 fejezetben bemutatott, az adaptív mérés vizsgálatára szolgáló eljárási mód, hanem a képletek alkalmazása különböző mérési keretek esetén, egyfajta illusztráció.

14. táblázat

Itemek hibája a kitöltők számának, a szintek számának és a teszt nehézségének (itemek megoldottsági valószínűségének) függvényében

K	p	N=10	N=20	N=30	N=50	N=100	N=200	N=300	N=400	N=500
2	0,1	1,451	1,026	0,838	0,649	0,459	0,324	0,265	0,229	0,205
2	0,2	1,054	0,745	0,609	0,471	0,333	0,236	0,193	0,167	0,149
2	0,3	0,886	0,626	0,511	0,396	0,280	0,198	0,162	0,140	0,125
2	0,4	0,791	0,559	0,456	0,354	0,250	0,177	0,144	0,125	0,112
2	0,5	0,730	0,516	0,422	0,327	0,231	0,163	0,133	0,116	0,103
3	0,1	1,762	1,246	1,017	0,788	0,557	0,394	0,322	0,279	0,249
3	0,2	1,268	0,896	0,732	0,567	0,401	0,284	0,232	0,200	0,179
3	0,3	1,054	0,745	0,609	0,471	0,333	0,236	0,193	0,167	0,149
3	0,4	0,930	0,658	0,537	0,416	0,294	0,208	0,170	0,147	0,132
3	0,5	0,849	0,600	0,490	0,380	0,268	0,190	0,155	0,134	0,120
4	0,1	2,026	1,432	1,169	0,906	0,641	0,453	0,370	0,320	0,286
4	0,2	1,451	1,026	0,838	0,649	0,459	0,324	0,265	0,229	0,205
4	0,3	1,201	0,849	0,693	0,537	0,380	0,269	0,219	0,190	0,170
4	0,4	1,054	0,745	0,609	0,471	0,333	0,236	0,193	0,167	0,149
4	0,5	0,956	0,676	0,552	0,428	0,302	0,214	0,175	0,151	0,135
5	0,1	2,259	1,597	1,304	1,010	0,714	0,505	0,412	0,357	0,319
5	0,2	1,614	1,141	0,932	0,722	0,510	0,361	0,295	0,255	0,228
5	0,3	1,332	0,942	0,769	0,596	0,421	0,298	0,243	0,211	0,188
5	0,4	1,166	0,824	0,673	0,521	0,369	0,261	0,213	0,184	0,165
5	0,5	1,054	0,745	0,609	0,471	0,333	0,236	0,193	0,167	0,149
6	0,1	2,470	1,747	1,426	1,105	0,781	0,552	0,451	0,391	0,349
6	0,2	1,762	1,246	1,017	0,788	0,557	0,394	0,322	0,279	0,249
6	0,3	1,451	1,026	0,838	0,649	0,459	0,324	0,265	0,229	0,205
6	0,4	1,268	0,896	0,732	0,567	0,401	0,284	0,232	0,200	0,179
6	0,5	1,144	0,809	0,661	0,512	0,362	0,256	0,209	0,181	0,162
7	0,1	2,665	1,884	1,539	1,192	0,843	0,596	0,487	0,421	0,377
7	0,2	1,898	1,342	1,096	0,849	0,600	0,424	0,347	0,300	0,268
7	0,3	1,561	1,104	0,901	0,698	0,494	0,349	0,285	0,247	0,221
7	0,4	1,362	0,963	0,787	0,609	0,431	0,305	0,249	0,215	0,193
7	0,5	1,228	0,868	0,709	0,549	0,388	0,275	0,224	0,194	0,174
8	0,1	2,846	2,013	1,643	1,273	0,900	0,636	0,520	0,450	0,403
8	0,2	2,026	1,432	1,169	0,906	0,641	0,453	0,370	0,320	0,286
8	0,3	1,665	1,177	0,961	0,744	0,526	0,372	0,304	0,263	0,235
8	0,4	1,451	1,026	0,838	0,649	0,459	0,324	0,265	0,229	0,205
8	0,5	1,306	0,924	0,754	0,584	0,413	0,292	0,239	0,207	0,185

Megjegyzés. Vastagítással jeleztük azokat a kombinációkat, ahol az item mérési hibája alatta marad a teljesítmény szórása 20 százalékának.

Mit láthatunk akkor, ha mindezt kiegészítjük az egyes tesztkitöltések hosszával? Egy 10 itemből álló teszt esetében lényegében nem nagyon tudjuk elérni, hogy a teljesítmény hibája elérje a teljesítmény szórásának 0,4-es vagy 0,5-ös minimális megbízhatósági szintjét (és ezt egyéb iránt az empirikus tapasztalatok is alátámasztják, ennyire rövid teljesítményt mérő tesztek általában nincsenek). Ezzel szemben 0,3-as nehézség esetében 30 itemmel már akár 20–30 kitöltővel is elérhető a teljesítmény 0,5-ös átlagos hibája. Igaz ez K = 3-ra, K = 5-re és K = 8 esetben is. Ez azt jelenti, hogy egy aránylag bonyolultabb teszt esetében, akár még 8 szintet megkülönböztetve is, a teljes

teszt megfelelő pontossága már 30–40 kitöltővel is elérhető, amennyiben megfelelő itembankkal rendelkezve adaptív tesztet tudunk összeállítani (15. táblázat).

15. táblázat

Teljesítmények hibája a kitöltők számának, a szintek számának, a teszt hosszának és a teszt nehézségének függvényében

K	L	p	N=10	N=20	N=30	N=50	N=100	N=200	N=300	N=400	N=500
3	10	0,1	1,054	1,524	1,807	2,052	2,244	2,341	2,374	2,391	2,401
3	10	0,2	0,791	1,021	1,17	1,302	1,407	1,461	1,479	1,489	1,494
3	10	0,3	0,69	0,835	0,932	1,021	1,092	1,129	1,141	1,147	1,151
3	10	0,4	0,645	0,746	0,816	0,881	0,933	0,961	0,97	0,975	0,978
3	10	0,5	0,632	0,707	0,76	0,809	0,85	0,872	0,879	0,882	0,885
3	30	0,1	2,527	0,864	0,609	0,782	1,052	1,207	1,26	1,287	1,303
3	30	0,2	1,512	0,582	**0,456**	0,54	0,68	0,764	0,793	0,808	0,817
3	30	0,3	1,121	**0,477**	**0,398**	**0,45**	0,541	0,597	0,617	0,627	0,633
3	30	0,4	0,914	**0,427**	**0,373**	**0,408**	**0,474**	0,515	0,529	0,537	0,541
3	30	0,5	0,792	**0,405**	**0,365**	**0,391**	**0,44**	**0,472**	**0,483**	**0,489**	**0,493**
3	60	0,1	4,319	1,772	0,961	**0,463**	0,551	0,74	0,811	0,848	0,87
3	60	0,2	2,547	1,061	0,6	**0,338**	**0,381**	**0,479**	0,517	0,537	0,549
3	60	0,3	1,857	0,787	**0,463**	**0,291**	**0,318**	**0,381**	**0,407**	**0,42**	**0,429**
3	60	0,4	1,487	0,642	**0,394**	**0,27**	**0,288**	**0,334**	**0,352**	**0,362**	**0,368**
3	60	0,5	1,259	0,557	**0,357**	**0,263**	**0,276**	**0,311**	**0,325**	**0,333**	**0,337**
5	10	0,1	1,054	1,652	1,997	2,292	2,52	2,636	2,675	2,695	2,707
5	10	0,2	0,791	1,11	1,305	1,476	1,61	1,678	1,701	1,712	1,719
5	10	0,3	0,69	0,907	1,045	1,168	1,264	1,314	1,331	1,339	1,344
5	10	0,4	0,645	0,81	0,917	1,014	1,091	1,131	1,144	1,151	1,155
5	10	0,5	0,632	0,765	0,854	0,936	1,001	1,034	1,046	1,052	1,055
5	30	0,1	2,899	0,934	0,609	0,832	1,164	1,35	1,413	1,445	1,464
5	30	0,2	1,795	0,63	**0,456**	0,573	0,76	0,867	0,905	0,923	0,935
5	30	0,3	1,371	0,516	**0,398**	**0,477**	0,608	0,685	0,712	0,726	0,734
5	30	0,4	1,151	**0,462**	**0,373**	**0,432**	0,533	0,594	0,616	0,627	0,633
5	30	0,5	1,027	**0,437**	**0,365**	**0,413**	**0,496**	0,548	0,566	0,575	0,58
5	60	0,1	4,986	2,033	1,083	**0,474**	0,586	0,818	0,904	0,948	0,974
5	60	0,2	3,059	1,259	0,689	**0,345**	**0,404**	0,534	0,584	0,61	0,625
5	60	0,3	2,316	0,962	0,539	**0,296**	**0,336**	**0,428**	**0,463**	**0,482**	**0,493**
5	60	0,4	1,927	0,808	**0,464**	**0,274**	**0,305**	**0,375**	**0,403**	**0,418**	**0,427**
5	60	0,5	1,702	0,721	**0,424**	**0,267**	**0,291**	**0,35**	**0,373**	**0,385**	**0,393**
8	10	0,1	1,054	1,775	2,176	2,516	2,778	2,911	2,955	2,977	2,991
8	10	0,2	0,791	1,196	1,435	1,64	1,799	1,88	1,908	1,921	1,929
8	10	0,3	0,69	0,979	1,155	1,308	1,428	1,489	1,509	1,52	1,526
8	10	0,4	0,645	0,874	1,017	1,143	1,242	1,292	1,309	1,318	1,323
8	10	0,5	0,632	0,826	0,949	1,059	1,146	1,191	1,206	1,213	1,218
8	30	0,1	3,24	1,002	0,609	0,881	1,269	1,483	1,556	1,593	1,615
8	30	0,2	2,052	0,677	**0,456**	0,607	0,836	0,965	1,009	1,032	1,045
8	30	0,3	1,598	0,555	**0,398**	0,504	0,672	0,769	0,802	0,819	0,829
8	30	0,4	1,365	**0,497**	**0,373**	**0,456**	0,592	0,671	0,699	0,713	0,721
8	30	0,5	1,238	**0,47**	**0,365**	**0,435**	0,552	0,621	0,646	0,658	0,665
8	60	0,1	5,594	2,272	1,196	**0,485**	0,62	0,892	0,991	1,041	1,072
8	60	0,2	3,522	1,439	0,772	**0,352**	**0,428**	0,588	0,647	0,678	0,697
8	60	0,3	2,726	1,121	0,611	**0,302**	**0,356**	**0,473**	0,517	0,54	0,554
8	60	0,4	2,316	0,958	0,53	**0,279**	**0,321**	**0,416**	**0,453**	**0,472**	**0,483**
8	60	0,5	2,09	0,869	**0,489**	**0,271**	**0,307**	**0,389**	**0,42**	**0,437**	**0,447**

Megjegyzés. Vastagítással jeleztük azokat a kombinációkat, ahol a teljesítmény mérési hibája alatta marad a teljesítmény szórása 50 százalékának.

6.3.4. Két példa gyakorlati felhasználásra

Az OECD PISA (OECD, 2019a) és az Országos kompetenciamérés (Auxné Bánfi et al., 2014) jellemzően olyan felmérések, ahol N >> L, azaz lényegesen, nagyságrendekkel több kitöltő diák van, mint ahány item egy-egy felmérés során felhasználásra kerül egy tesztfüzetben. Jellemzően egy tesztfüzet egy-egy területen 50–60 itemet tartalmaz, míg a kitöltők száma több ezres vagy tízezres nagyságrendű.

Mindkét teszt esetében próbateszteket szerveznek (e próbateszteken mérik be a későbbiekben használatra kerülő itemeket), ami azt jelenti, hogy az itembank, ami rendelkezésre áll nagyságrendileg 20–30-szorosa egy-egy évben a végül felhasználásra kerülő itemeknek. Ez egyben azt is jelenti, hogy a korábbi évek itemjeivel együtt olyan gazdag feladatbank áll rendelkezésre, hogy akár adaptív módon is könnyen lehet mérést szervezni az itemek kimerülésének, korrumpálódásának kockázata nélkül (Cizek & Wollack, 2016).

A szimulációs eredmények megmutatták, hogy 200–300 kitöltő esetében érdemi különbségek a szintek (K = 4 és felette) és a megoldottságok (p = 0,3 felett) között nincsenek, lényegében hasonló működéseket tapasztalunk. Az alábbi megkötések tehát gyakorlati szempontból életszerűek.

1) K = 5: iskolai környezetben az 1 és 5 közötti osztályzatok használata megszokott, ezeket mind a pedagógusok, mind a diákok megfelelő módon tudják értelmezni.
2) N legyen 100 és 500 közötti rögzített érték. Jellemzően az OECD PISA és a Kompetenciamérés a próbamérések során az itemeket nagyjából ennyi diákkal tölteti ki. Az általános tapasztalatok alapján N = 300 megfelelő értéknek mutatkozik.
3) p = 0,3, p = 0,4 és p = 0,5 esetekkel dolgozzunk, ennél nehezebb vagy könnyebb teszteket jellemzően nem szokás íratni – legalábbis felmérés jelleggel azok a tesztek, amiket mindenki megold, illetve senki sem tud rajta érdemben jól teljesíteni, tömegesen nem alkalmazottak.

A fenti megkötések azért lehetnek érdekesek, mert jelen elemzés fókusza alapvetően nem az, hogy 100 vagy 500 kitöltőre van szükség. A mostani infrastruktúra mellett 300 diákkal egy teljesítményt mérő tesztet kitöltetni érdemi költségekkel nem jár. A K = 5 megkötés nem érdemi megkötés egy iskolai rendszerben. A tesztek nehézségének 30–70% közötti rögzítése szintén kellően bő keretet szolgáltathat tesztek összeállításához.

Megfigyelhető (16. táblázat), hogy egy nehezebb tesztnél (30%-os megoldottság) már 54 kérdésnél elérhető a megfelelő megbízhatósági szint. Könnyebb tesztnél (40%-os megoldottság) 43 itemes teszt mutatja az alsó határt, míg egy alapvetően kiegyensúlyozott, 50%-os megoldási szintre beállított teszt esetében 38 kérdésből álló tesztekkel már elfogadható megbízhatósági szintet érhetünk el.

16. táblázat

Könnyebb, ötfokozatú (a) és nehezebb, két fokozatú (b) tesztek minimális teszthossza

Teszt várható hossza	Megoldás valószínűsége		
	0,3	0,4	0,5
30	0,712	0,616	0,566
31	0,699	0,604	0,555
32	0,686	0,593	0,545
33	0,673	0,582	0,535
34	0,661	0,572	0,526
35	0,650	0,563	0,517
36	0,639	0,553	0,509
37	0,629	0,544	0,501
38	0,618	0,536	**0,493**
39	0,609	0,527	**0,485**
40	0,599	0,519	**0,478**
41	0,590	0,512	**0,471**
42	0,582	0,504	**0,465**
43	0,573	**0,497**	**0,458**
44	0,565	**0,490**	**0,452**
45	0,557	**0,484**	**0,446**
46	0,550	**0,477**	**0,440**
47	0,542	**0,471**	**0,434**
48	0,535	**0,465**	**0,429**
49	0,528	**0,459**	**0,423**
50	0,522	**0,453**	**0,418**
51	0,515	**0,447**	**0,413**
52	0,509	**0,442**	**0,408**
53	0,502	**0,437**	**0,403**
54	**0,496**	**0,432**	**0,399**
55	**0,491**	**0,427**	**0,394**
56	**0,485**	**0,422**	**0,390**
57	**0,479**	**0,417**	**0,385**
58	**0,474**	**0,412**	**0,381**
59	**0,469**	**0,408**	**0,377**
60	**0,463**	**0,403**	**0,373**

a) K = 5, N = 300.

Teszt várható hossza	Megoldás valószínűsége		
	0,2	0,25	0,3
30	0,726	0,625	0,557
31	0,713	0,614	0,547
32	0,700	0,603	0,538
33	0,688	0,593	0,529
34	0,677	0,584	0,520
35	0,666	0,575	0,512
36	0,656	0,566	0,504
37	0,646	0,557	**0,497**
38	0,637	0,549	**0,490**
39	0,628	0,541	**0,483**
40	0,619	0,534	**0,476**
41	0,610	0,527	**0,470**
42	0,602	0,520	**0,464**
43	0,594	0,513	**0,458**
44	0,587	0,506	**0,452**
45	0,579	0,500	**0,446**
46	0,572	**0,494**	**0,441**
47	0,565	**0,488**	**0,436**
48	0,558	**0,482**	**0,431**
49	0,552	**0,477**	**0,426**
50	0,546	**0,471**	**0,421**
51	0,539	**0,466**	**0,416**
52	0,533	**0,461**	**0,412**
53	0,528	**0,456**	**0,407**
54	0,522	**0,451**	**0,403**
55	0,517	**0,447**	**0,399**
56	0,511	**0,442**	**0,395**
57	0,506	**0,438**	**0,391**
58	0,501	**0,433**	**0,387**
59	**0,496**	**0,429**	**0,384**
60	**0,491**	**0,425**	**0,380**

b) K = 2, N = 300.

Megjegyzés. A 0,5-ös hibahatár elérését vastagítással jeleztük.

K = képességszintek száma, N = kitöltők várható száma.

A 16. táblázatból látható, illetve a PISA és az OKM általános tapasztalatai azt mutatják, hogy lineáris tesztek esetében akár 50–70 kérdés is megtalálható a tesztfüzetekben. Adaptív módon kevesebb, akár 40 kérdéssel már egy szóráshoz képest 40% körüli hibával rendelkező teszt is összeállítható. Egy ilyen adaptív teszt az alábbi előnyökkel rendelkezik:

1) Lényegesen kevesebb itemmel, kevesebb idő alatt tudunk felmérést készíteni.
2) A diákok a nekik megfelelő szintű feladatokat kapják (Kingsbury, 2009), tehát nem unják el a feladatokat, jellemzően mindenki a számára még kihívást jelentő itemekkel dolgozik.
3) Ha van a felmérésből fennmaradó idő, akkor a próbafelmérés során kipróbálandó itemek azok, amelyeket a diákok a fennmaradó időben oldhatnak – így a következő mérés próbaidőszaka az előző időszakkal összevonható, párhuzamosan vezényelhető.

A fenti előnyök fenntartása mellett egy második példát is bemutatok. Elsősorban felsőoktatásban vagy szakképzései területeken fordulnak elő olyan tesztek, melyek szintén adaptívvá tehetők és valójában $K = 2$ szintet követelnek meg (teljesített vagy nem teljesített). Ez esetben nem feltétlenül osztályzatokat képezünk, hanem egy elvárt szint teljesítését tűzzük ki célként.

Ilyen esetben jellemzően nehezebbek szoktak lenni a teljesítések, tehát $p = 0{,}2$, $p = 0{,}25$ és $p = 0{,}3$ eseteket mutatom be (azaz, a teljesítéshez egy 20%-os sikerességi határt veszek, mind legnehezebb teljesítési szintet). Ez értelemszerűen úgy is interpretálható, hogy a vizsga teljesítéséhez legalább 80%-os teljesítményre van szükség. A megoldók száma továbbra is legyen $N = 300$ főben rögzítve, mely akár felsőoktatási, akár szakképzési keretek között nem életidegen feltételezés. Bár a rendszer szigorúbb, azonban itt is 0,5-ös megbízhatóságot fogok minimum határként megadni – tehát vizsgázónként hasonlóan pontos becslést szeretnék az adaptív teszttől átlagosan elvárni.

Ebben az esetben látható (16. táblázat), hogy 30%-os, némileg megengedőbb teljesítési szint esetében már 37 kérdésnél elérhető a 0,5-ös megbízhatóság szint. Kicsit szigorítva, $p = 0{,}25$-nél ehhez minimum 46 itemre van szükség, illetve $p = 0{,}2$ esetében minimum 59 item lesz az alsó határ.

Ez azt is jelenti, hogy adaptív módon egy teljesített/nem teljesített rendszer működtetésénél 60 item hosszú teszt már elegendő lehet adaptív módot működtetve annak kiderítésére, hogy az adott vizsgázó valóban megfelelő szinten helyezkedik-e el. Abban

az esetben ugyanis, ha 60 item segítségével egy nehezebb tesztet töltetünk ki, a teljesítményt mérő modellek segítségével a megfelelő képességpontja számítható, abból pedig láthatóvá válik, hogy eléri-e a számunkra elfogadható szintet vagy sem.

6.3.5. Speciális eset: kiegyensúlyozott megoldottságú itemek (p = ½)[43]

Az előző példákhoz képest további egyszerűsítést remélek attól a szituációtól amely a legtöbb adaptív teszt esetében fennáll. Vizsgálatomban feltételezem, hogy azt a következő itemet választom, mely esetén a megoldás valószínűsége éppen 0,5 (azaz 50%). Ilyen kiválasztási eljárások pl. az aktuális képességpontban a legnagyobb Fisher-információval rendelkező item (MFI) vagy az aktuális becsült pontszámhoz legközelebbi nehézségű item (bOpt) (ld. 2.3.1 és 6.4 fejezetekben) kiválasztásának módszere. Ebben a speciális esetben a cél az, hogy dichotóm kérdésekkel egydimenziós jelenség vizsgálata esetén zárt formulával legyen meghatározható, hogy hány item segítségével lehet adott lineáris teszthez hasonló becslési pontosságot elérni – a teljes teszt tekintetében és egyénekre vonatkoztatva. További cél zárt formulát adni arra is, hogy az itemek megfelelő pontosságú próbaméréséhez meghatározzuk, hány kitöltőre van szükség. Ehhez minden esetben ideális körülményeket feltételezek, azaz eredményeim a hasonló keretben vizsgált empirikus eredmények vagy való életbeli fejlesztések kemény korlátai lesznek.

6.3.6. Próbamérés szükséges nagysága

A standard hiba akár item, akár kitöltők esetében azt a jelentést hordozza, hogy úgy általában milyen mérési pontossággal bírunk. Ez nem azt jelenti, hogy nem lehetnek ennél sokkal pontosabban értékelt kitöltők vagy itemek – de általában azt tudjuk, hogy ezt a szintet teljesíteni fogjuk. Ha például 8 szinten akarunk mérni, akkor egy adott itembankba kerülő itemek jellemzően ezt a teljes spektrumot fedni tudják – így minden kitöltő esetében várhatóan kellő pontosságot tudunk elérni. Ehhez az alsó és felső szinteken mérő, a közepes tartományban található itemekhez képest ugyanolyan pontosságú feladatokra van szükség (itt is teljesíteni akarjuk a hibahatárokat), azonban az alacsony és magas képességű kitöltők száma relatíve alacsonyabb.

[43] A alfejezet a *Quality Assurance in Education* folyóiratban megjelent cikk (T. Kárász et al., 2023) alapján készült.

Az itemek megfelelő pontosságú beméréséhez tételezzük fel a következő, idealisztikus szituációt. Vegyünk egy mérési rendszert, melyben a képességskálát K egyenlő hosszú részre osztják fel, és az itemeket nehézségük alapján nehézségi szintre is besorolják, ez a szokásos módon (OECD, 2019a) meghatározza a képességpontok és képességszintek beosztását is. Vegyünk továbbá egy próbamérést, ahol a mérés kezdete előtt ismert a kitöltők képességpontja. Vegyük észre, hogy ez nem lehetetlen feltétel olyan mérések esetében, ahol számítógépes a megvalósítás, azonnali a válaszok kiértékelése, és a méréshez kapcsolódik a próbamérés. Ha a mérés eredményeképpen ismert a kitöltők képessége, akkor lehetséges, hogy az egyes tesztfüzeteket, vagyis próbaitem-csomagokat ne véletlenszerűen osszuk ki, hanem egyenletesen, azaz olyan tanulói csoportoknak, melyek azonos számban képviselik az adott itemszintet. Egyszerű véletlen mintavétel esetén az itemet kitöltők képességpontjai várhatóan normális eloszlást követnek, ezért a képességskála szélein elhelyezkedő itemeket pontatlanabbul lehet kalibrálni. Egyenletes mintavétel esetén garantálható, hogy végül bármely nehézségi szintre eső itemet N/K releváns válasz alapján értékeljünk, ahol N a teljes próbamérés mintanagysága, azaz mindazon kitöltők száma, akik találkoztak az itemmel. Relevánson azt értjük, hogy hasonlóan a képességpontokhoz, a legtöbb információt azok válaszaiból nyerhetjük, akiknek képességpontja a legközelebb van az item nehézségéhez. Ez ugyanaz a nehézségi szint, amelyen a későbbi mérések az itemet alkalmazni, választani fogják. A hasonló számú kitöltöttség eredménye, hogy minden itemet hasonló mérési pontossággal tudunk jellemezni. Induló feltételeink mellett, azaz egydimenziós jelenséget független dichotóm itemekkel mérő teszt esetében és számítógépes megvalósítás mellett tehetünk néhány további megállapítást.

1) Item-csomagok helyett egy-egy item beméréséreis törekedhetünk, ebben az esetben figyelve az egy kitöltőhöz kerülő itemek tartalmi hasonlóságra.
2) Amennyiben rendelkezünk valamilyen előzetes információval vagy szakértői véleménnyel az item nehézségéről, megtehetjük, hogy nem az összes, csak néhány nehézségi szinten végezzük a vizsgálatát. Ebben az esetben N továbbra is a mérésben részt vevők száma, K azonban azon szintek száma, melyeken kipróbáltuk az itemet. Ebben az értelemben azt a küszöbértéket keressük, hogy nehézségi szintenként hány kitöltőre van szükség az item jellemzőinek kellően pontos meghatározásához, a mintanagyság pedig ennek szorzata a kipróbált nehézségi szintek számával.

3) Egyenletes választás esetén a populációs középérték közelében lényegesen nagyobb számú kitöltői csoportot alakíthatunk ki, ami több közepes item bemérésére ad lehetőséget (feltéve, hogy nem minden szinten próbáljuk ki a feladatokat). Mivel az itemek kitettségének korlátozása miatt az itembankban több itemre van szükség a középső tartományban, ez több közepes nehézségű item bemérésére ad lehetőséget.
4) Szélsőséges esetben elegendő lehet egyetlen nehézségi szintet kiválasztani a kipróbáláshoz. Ez éppen az a szint, amelyiken az item a nehézsége alapján van. Ez egyben azt is jelenti, hogy a többi szinten felvett válasz a képességskála nehézségétől távolabbi részén szolgáltat információt.

Mivel a releváns képességszinten az item nehézsége és a kitöltők képessége közel van egymáshoz, ezért az item megoldási valószínűsége a Rasch-modell tulajdonságaiból következően várhatóan 1/2 (50%) körül van. A fenti (1) egyenletet (115. oldal) használva egy közepes item megoldottságot (50%) feltételezve az alábbi eredményhez jutunk:

(3) $$N = \left\lceil \frac{1}{S_1^2} \frac{4K^2}{2K-1} \right\rceil,$$

ahol S_1 jelöli az adott item hibáját, a korábban $SE(d_i)$-vel jelölt mennyiséget.

Ez alapján a szükséges esetszám dinamikáját két oldalról tudjuk vizsgálni:

1) Minél pontosabb itemjellemzőket szeretnénk elérni (minél kisebb az elérendő standard hiba), várhatóan annál több kitöltőre van szükségünk.
2) Minél több szinten vizsgáljuk az adott item viselkedését, annál több kitöltőre van szükség. (Ekkor *K* itemszinten szerzünk egyformán pontos információt az item működéséről).

Ezek nem meglepő állítások, azonban a fenti formula ismeretében a szintek száma és az elvárt hibahatár alapján meg tudjuk határozni a minimálisan szükséges mintanagyságot. Ez természetesen nem jelenti azt, hogy az item (tartalmi keret vagy műveleti szint szerint) illeszkedik a tesztbe. Csak az biztosított, hogy egyébként szakmailag jónak látszó item jellemzőinek mérési pontossága megfelelő.

6.3.7. Teljesítménybecslés adott hibája mellett szükséges teszthossz

Maga a mérés esetében a részt vevő *N* kitöltőt szeretnénk *K* képességszinten elhelyezni (valamint lesz egy legalsó szint alatti képességkategória is) adaptív mérés segítségével. Tegyük fel, hogy ehhez a rendszer gazdag itembankkal rendelkezik, azaz bármely becsült

képességponthoz van kellő számú ugyanolyan nehézségű item. Célunk meghatározni, hogyha előre rögzítjük azt az általános mérési pontosságot, amit a teszteredmények esetében szükségesnek tartunk, akkor várhatóan hány item hosszúságúak lesznek az egyébként eltérő tesztutak.

Az adaptív tesztek esetében az úgynevezett megállítási kritériumok (*stopping rules*) szerves részét képezik a mérési rendszernek (Stafford et al., 2019). Egy megállítási kritérium jellemzően tartalmaz valamilyen alsó és felső korlátot az itemek számára és/vagy az eltelt időre vonatkozóan. A mérés céljától függően valamilyen becslési hibával kapcsolatos határ elérése vagy valamilyen kategóriákba történő besorolás bizonyossága is része lehet, előbbi éppen megfelel jelenlegi vizsgálatom céljának.

A *b* kitöltő teljesítményének hibájára vonatkozó (2) általános formula (115. oldal):

$$S_2 = SE(b) = \sqrt{\frac{1}{Lp(1-p)}}\sqrt{1 + \frac{\left(\frac{L}{L-1}\right)\left(\frac{N-L}{N}\right)^2\left(\frac{K-p}{p}\right)}{2{,}89}},$$

ahol L a teszt hossza a b becsült képességfejlettségű kitöltő esetében, p pedig az egyes itemek megoldási valószínűsége. A levezetésben kihasználtam, hogy adaptív mérés és ideális feladatbank esetén szabályozhatjuk a megoldás valószínűségét, a most vizsgált kiegyensúlyozott esetben ez 1/2 lesz.

Felmerülhet a kérdés, hogy egy adott vizsgálati alany hibájának meghatározása miért függ az összes kitöltő számától (N). Az IRT modellekben, illetve a modern tesztelméletben jellemzően az itemek nehézsége és a kitöltők teljesítményének meghatározása iteratív folyamat. Ez azt jelenti, hogy az itemek nehézsége és a kitöltők képességei azonos skálára kerülnek, tehát egyszerre kell beállítani az adott itemek nehézségeit, illetve a kitöltők képességeit. Wright formulái ezen a modellen alapulnak. Adaptív mérések esetén, az itemparaméterek ismerete felveti a lehetőségét ezen tag elhagyásának, mindazonáltal nagymintás mérések esetében N és L nagyságrendje alapján az $(N - L) / N$ tag nagyságrendje 1-hez közeli.

A fentiek alapján p = 1/2 behelyettesítésével L értékére az alábbi levezetést írhatjuk fel.

$$S_2 = \sqrt{\frac{4}{L}}\sqrt{1 + \frac{\left(\frac{L}{L-1}\right)\left(\frac{N-L}{N}\right)^2\left(\frac{K-p}{p}\right)}{2{,}89}}$$

$$S_2 = 2\sqrt{\frac{1}{L} + \frac{\left(\frac{1}{L-1}\right)\left(\frac{N-L}{N}\right)^2 (2K-1)}{2{,}89}}$$

$$S_2 = 2\sqrt{\frac{2{,}89 + \left(\frac{L}{L-1}\right)\left(\frac{N-L}{N}\right)^2 (2K-1)}{2{,}89L}}$$

$$\frac{S_2^2}{4} = \frac{2{,}89 + \left(\frac{L}{L-1}\right)\left(\frac{N-L}{N}\right)^2 (2K-1)}{2{,}89L}$$

$$\frac{2{,}89S_2^2}{4} L = 2{,}89 + \left(\frac{L}{L-1}\right)\left(\frac{N-L}{N}\right)^2 (2K-1)$$

$$\frac{2{,}89S_2^2}{4(2K-1)} L = \frac{2{,}89}{(2K-1)} + \frac{L}{L-1}\left(\frac{N-L}{N}\right)^2$$

Az *L* teszthossz szempontjából konstans tagokra alkalmazzuk a következő jelöléseket:

$$D = \frac{2{,}89}{(2K-1)}$$

$$A = \frac{2{,}89S_2^2}{4(2K-1)} = \frac{S_2^2}{4} D$$

Ezen helyettesítések mellett a fenti formula az alábbi módon írható át:

$$AL = D + \frac{L}{L-1}\left(\frac{N-L}{N}\right)^2$$

Innen átalakítások után adódik, hogy

(4) $$0 = L^3 - L^2(AN^2 + 2N) + L(N^2 + DN^2 + AN^2) - DN^2$$

E harmadfokú egyenlet megoldása(i) adják meg azokat a teszthosszokat, melyek adott beállítások mellett a megfelelő mérési hibára vezetnek, ezek között találjuk meg a minimálisan szükséges teszt hosszát.

A képességpont becslésének hibája esetében fontos fejben tartanunk, hogy az S_2 hibatag nem jelenti az összes kitöltő mérési hibájának egyedi pontosságát. Helyette minden képességszinten egy átlagos hibanagyságot szeretnénk biztosítani – tehát a legalsó és legfelső szinteken is a középső szinteken elért pontosságot. Továbbra is lehetnek kitöltők, akiket ennél akár sokkal pontosabban tudunk mérni, illetve olykor pontatlanul, nagyobb hibával terhelten. Szintén lehetnek kitöltők, akik az átlagos minimális teszthossznál rövidebb vagy hosszabb tesztet töltenek ki.

6.3.8. Példák

Mindkét fenti eredmény példákkal illusztrálható. Először azt nézzük, hogy egyes itemek paramétereinek bemérésekor *K* nehézségi szint mellett hány ismert képességfejlettségű kitöltőre van szükség. Ezt nevezhetjük a próbaméréshez szükséges kitöltők számának is. Másodszor pedig azt, hogy ha ismert paraméterű itemekkel rendelkezünk, akkor ideális itembank és adaptív mérés mellett várhatóan milyen hosszú tesztekre lehet számítani.

Az itemek jellemzőinek meghatározása esetén az itemek elvárt mérési hibáját (S_1) 0,1 és 0,5 között vizsgálom, tizedenkénti lépésközzel. A nehézségi szintek számát (*K*) 2 és 8 közötti értékekben határozom meg. A $K = 2$ a megfelelt/nem felelt meg kategóriákat jelentheti, míg a magasabb finomságú skálák a nagymintás tanulói felmérések szintjeihez igazodó értékek. A szintek száma és a feladatok átlagos hibája alapján az alábbi létszámok adódnak (17. táblázat).

Megfigyelhető, hogy a szintek számával – tetszőleges sort nézve, adott átlagos hiba mellett – a kitöltők száma növekszik, ami megfelel az elvárásoknak: minél több szinten, minél inkább nagyobb finomsággal szeretnénk az itemeket bemérni, annál több kitöltőtől kell válaszokhoz jutni. Ez a mintánk *K* szintre darabolásának következménye.

Ezzel párhuzamosan a szintek számát rögzítve és a hibát növelve csökken (illetve a hibát csökkentve növekszik) a szükséges kitöltők száma. Ez is elfogadható, hiszen a kitöltők száma esetében jellemzően azt mondhatjuk, hogy minél pontosabban szeretnénk egy adott item jellemzőit mérni (minél alacsonyabb hibával dolgozunk), várhatóan annál több kitöltőre van szükségünk. Továbbá a nagymintás méréseknél egy-egy itemet jellemzően 300–500 kitöltő lát a próbák során (véletlen mintavétel mellett), és 0,2–0,3 közötti hibák szoktak adódni (Robitzsch & Lüdtke, 2019), ilyen értelemben az egyenlet eredményei egybe esnek az empirikus tapasztalatokkal.

17. táblázat

Próbaméréshez szükséges kitöltők száma a szintek és az itemek átlagos standard hibájának függvényében

Itemek mérési hibája (S_1)	Szintek száma (K)						
	2	3	4	5	6	7	8
0,1	534	721	915	1112	1310	1508	1707
0,2	134	181	229	278	328	377	427
0,3	60	81	102	124	146	168	190
0,4	34	46	58	70	82	95	107
0,5	22	29	37	45	53	61	69

Most vizsgáljuk meg azt, hogy a tesztek várható hossza (L) miként alakul akkor, ha a képességbecslés hibáját (S_2) és a szintek számát (K) rögzítjük – valamint a kitöltők számát (N) realisztikusra állítjuk. Az elvárt hibahatárt 0,4, 0,5 és 0,6, a szintek számát 2, 5 és 8 értékben határozom meg. A kitöltés mintanagyságát (N) egy nagyobb egyetemi évfolyam méretében (200–300 fő, jelen esetben 250 főben rögzítem), egy közepes/kisebb ország nemzetközi felmérésben történő részvételében (4000–5000 fő, elemzésemben 4500 fő) és egy országos mérés egy évfolyamának létszámában (80 000–110 000 fő, elemzésemben 10 0000 fő) határozom meg. A várható teszthosszra (L) kapott eredményeket, azaz a harmadfokú egyenlet valós gyökét a 18. táblázat foglalja össze.

Az eredmények közül kiemelendő a 0,4-es (tehát legkisebb hiba, leginkább pontos becslés), 250 kitöltő és 8 képességszint esetében mutatott hiányzó eredmény. Ez azt jelenti, hogy alacsony kitöltési számnál, nagy finomságú mérés esetében nem tudunk tesztet alkotni, ami reális eredmény.

Az országos/nemzetközi mérések a 0,5 körüli hibaértékeket szoktak elérni (Robitzsch & Lüdtke, 2019), ami egy átlagos szintnek mondható. Ennél pontosabban vagy pontatlanabbul mért tesztalanyok is vannak, de összességében ezt az átlagos szintet szeretnénk adaptív méréssel minden értékelési szinten (K = 8 esetében is) elérni. Ebben az esetben szintén a legfinomabb kategorizálás esetében adódnak a leghosszabb tesztek. A magyarországi 6., 8. és 10. évfolyamon teljeskörű OKM esetében (évfolyamonként 100.000 kitöltő, átlagos hiba 0,5 és 0,6 közötti, 7+1 képességszint (Auxné Bánfi et al., 2014)) a 2021-ig papír-ceruza alapú, 2022-től számítógép alapú lineáris teszt nagyjából 60 kérdést tartalmaz. Egy ennek megfelelő, kisebb finomságú, de még mindig részletes képességfelosztású adaptív teszt esetében rövidebb (41–58 hosszú teszt) elegendő.

Egy közepes méretű egyetemi évfolyam esetében az egyetemi rendszerben átlagos (0,5–0,6-es) megbízhatóság mellett 17–25 kérdést tartalmazó adaptív teszt segítségével már a vizsgák (megfelelt/nem felelt meg) értékelés elérhető. Hagyományos, 5 fokozatú skálát alkalmazva 40–60 kérdéssel minden képességszinten tartható ez a hibahatárt.

18. táblázat

Tesztek várható hossza a bemért szintek, a kitöltők számának és a képességbecslés elvárt standard hibájának függvényében

Egyéni mérés hibája (S_2)	Létszám (N)	Szintek száma (K)		
		2	5	8
0,4	250	40	114	NA
	4500	35	68	92
	100000	35	68	90
0,5	250	25	58	94
	4500	24	44	59
	100000	23	44	58
0,6	250	17	37	54
	4500	17	31	41
	100000	17	31	41

6.3.9. Összegzés

Megállapítható, hogy még akár igen nehéz teszteket, a diákok számára kihívást jelentő kérdéseket alkalmazva is 60 kérdést összeállítva megfelelő pontosság érhető el ahhoz, hogy a megoldott feladatokból kiszámítva a diák képességpontját, általánosságban elfogadható értékelést tudjunk biztosítani.

Ezzel szemben jellemző iskolai körülményeket szimulálva a tesztekhez (K = 5, tehát 5 fokozatú értékelést alkalmazva, p = 0,4, illetve p = 0,5, azaz átlagos tesztnehézséget feltételezve) ez az itemszám lényegesen kevesebb, 40–50 feladatnál áll meg. Ez lényegében megegyezik azokkal a mutatókkal, melyekkel a nemzetközi adaptív mérések (pl. OECD, 2019a) esetében általánosságban találkozunk.

Továbbá ez azt is jelenti, hogy egy adaptív teszt esetében a mostani mérési idő jelentősen csökkenthető (figyelembe véve, hogy a fentebb jelzett lineáris tesztek 50–60 itemet használnak tesztfüzetenként Auxné Bánfi et al., 2014), így az a cél, hogy a kifáradást elkerüljük, a diákok érdeklődését folyamatosan fenntartsuk, általánosságban is teljesülhet olyan tesztekkel, melyek rövidebbek és folyamatosan olyan kérdéseket adnak

a diákoknak, amik a tudásszintjüknek éppen megfelelnek. A fennmaradó idő a próbaitemeik bemérésére fordítható. Ez azt is jelenti, hogy egy-egy időszakban nem kell kétszeres logisztikát alkalmazni (próbaméréseket tartani az itemek tesztelésére), hanem az amúgy is mérésre fordított időben lehet a tesztelést megvalósítani. Tehetjük ezt akár úgy is, hogy a valós tesztitemek közé véletlenszerű helyekre tesszük a bemérendő próbafeladatokat (Cizek & Wollack, 2016). Ekkor egy adott diák számára nem rövidül jelentősen a tesztre fordított idő a lineáris teszthez képest, azonban nem szükséges külön próbamérésen részt venni, ami a tanulók körülbelül 3,5 százalékát érinti minden mérési évben.

Az elemzésben alkalmazott formula segítségével adaptív mérés előkészítési szakaszában tervezhető, hogy az egyes itemek, illetve vizsgálati személyek megfelelő pontosságú felmérése milyen minimális mintanagysággal és teszthosszal érhető el. Ez a megoldás eddig hiányzó hidat teremt az elmélet és a gyakorlat, azon belül is a szimulációk között. Mivel a formula levezetéséhez idealisztikus feltételek meglétét feltételeztem, ezért ezek a minimumok elmaradhatnak egy valóságos mérésben szükséges számoktól.

6.4. Lehetséges adaptív stratégiák összehasonlítása pontosság és megbízhatóság alapján – Szimulációs eredmények

Az eddigi eredmények alapján a papír ceruza mérés itemei jól adaptálhatók lehetnek számítógépes tesztelési környezetbe, azaz az itemek paraméterei nem térnek el jelentősen a papír ceruza mérés paramétereitől (ld. 6.1 fejezet). A nyílt itemek vizsgálatára vonatkozó kutatás eredményei alapján (ld. 6.2 fejezet) az azonnali kiértékelésre alkalmas zárt itemek elegendőek lehetnek a képességskála elég nagy részén a pontos képességbecsléséhez. Ugyanakkor a képesség skála szélein szükség lehet a nyílt itemek hasonló nehézségű itemekkel történő pótlására, esetleg innovatív itemtípusok kipróbálására. Elméleti levezetésem alapján (ld. 6.3 fejezet) a legegyszerűbb Rasch-modell esetén is elérhető adaptív méréssel a lineáris teszttel azonos megbízhatóságú becslési pontosság. Emellett a teszt várható hossza valamivel kisebb lehet, mint a lineáris teszt hossza. Ezek szerint az OKM tervezési fázisában a korábbi itemek paraméterei és a tanulói képességbecslések jó előrejelzői a számítógépes vagy az adaptív mérés adatainak, tehát szimulációs vizsgálatokban megfelelő minőségű eredményre juthatunk.

Kutatásom utolsó fázisában szimulációs vizsgálatokat végeztem, hogy összehasonlítsam néhány képességbecslési és itemkiválasztási módszer hatékonyságát. A

vizsgálat keretéül az OKM szolgált. Kutatási kérdésem arra irányult, hogy 1) két eltérő mérési cél esetében mely módszerek alkalmasabbak, ha gyorsabb vagy pontosabb tesztelést szeretnénk elérni, 2) az eddigi mérések itemei milyen adaptív mérési eredményeket prognosztizálnak.

Mindkét esetben hibrid szimulációt futtattam abban az értelemben, hogy a 2008–2019. évi mérések jól működő dichotóm itemei (625 item), és azok háromparaméteres modell szerint számított paraméterei alkották az itembankot[44]. A szimulációk így előzetes információt szolgáltatnak arról, hogy az évek során felhalmozott itemek digitalizált változatai megfelelők lehetnek-e egy adaptív méréshez. A tanulók elméleti képességfejlettségét a képességskála finom felosztása adta, 800 és 2200 pont közötti 50 pontos lépésközzel. Minden ponton kétszáz mérést szimuláltam. A szimulált tesztek belépési értéke a 6. évfolyamos országos átlag, 1500 pont volt. A képességbecslés Bayes-változatai esetében 1500 pont átlagú és 200 pont szórású prior eloszlást határoztam meg. A vizsgált képességbecslési eljárások a maximum likelihood (ML) (Lord, 1980), a Bayes-modal (BM) (Birnbaum, 1969) és az expected a-posteriori (EAP) (Bock & Mislevy, 1982) eljárások voltak. Az itemkiválasztási eljárások közül a Maximum Fisher információ (MFI) (Birnbaum, 1968), a legközelebbi nehézség (bOpt) (Urry, 1970) és a legközelebbi maximális információ (thOpt) (Barrada et al., 2006) kritériumokat hasonlítottam össze.

A szimulációk során két megállítási kritériumot vizsgáltam. A megállítási kritérium minden esetben a mérés céljához és a technikai korlátokhoz igazodik. Technikai korlát lehet az, ha elfogynak a teszt során kiosztható itemek, vagy a teszt kitöltője eléri a mérésre maximálisan felhasználható időkeretet. A mérés célja szerinti megállítás lehet az, ha előre rögzített a kiosztandó itemek száma, ami minden tesztalanynak a lehető legpontosabb, a hasonló hosszúságú papír-ceruza tesztnél várhatóan pontosabb becslést eredményez (ld. 6.4.1 fejezet). Egy másik cél lehet, amikor a tesztkitöltés várható idejét minimalizáljuk, ekkor a teszt akkor ér véget, amikor a képességbecslés hibája egy bizonyos határ alá kerül (ld. 6.4.2 fejezet), vagy az aktuális képesség becslése az előző értéktől valamilyen küszöbértéknél kisebb mértékben különbözik, azaz a változás már minimális. Mindkét esetben valamilyen konkrét képességpont előállítása a cél. Lehetnek olyan mérési célok, amikor pontos képességbecslés helyett a kitöltőt valamilyen teljesítményszintre sorolják be. Az OKM esetében ez jellemzően másodlagos cél, ezen

[44] Az itemek paraméterei megtalálhatók az OKM *Feladatok és jellemzőik* köteteiben.

kívül a képességszintek nagy száma (7+1) és a szimulációt végző program korlátai miatt jelenlegi formában a vizsgálat nem lehetséges.

6.4.1. Rögzített teszthosszhoz tartozó hiba becslése

A megállítási kritérium kiválasztása mögött egy szervezési feladat, a próbaitemek bemérésének egy lehetséges módja áll. A próbaitemeket a papír-ceruza tesztek esetében a trendszámítást lehetővé tevő Core itemekkel együtt az OKM kiegészítő mérésében töltötték ki a tanulók. Ez évente megközelítőleg 350–400 iskolát, illetve osztályt érintett, a májusi mérést megelőző időszakban. A kiegészítő mérésben az iskolák önkéntes alapon vesznek részt, ezért minden esetben a kiválasztott osztály mellé két másik iskola, mint pótiskola, szintén kiválasztásra került. Abban az évben, amikor valamely iskolai osztály a nemzetközi mérések mintájába kerül, a terheket csökkentendő, nem került felkérésre a kiegészítő mérésben.

A számítógépes mérés bevezetésével a kiegészítő mérés megszűnt. A szükséges próbaitemek bemérése őszi méréssel történt, a korábbi megbízható itemek híd itemei a mérésbe kerültek bemérésbe. A kényszerű őszi próbamérés hosszabb távon nem fenntartható, nem is célravezető, a tavaszi mérési időszak pedig rendkívül megterhelő, ezért egy lehetséges megoldás lehet, ha a próbaitemeket a főmérésbe „rejtve” töltik ki a tanulók. Ezek az itemek nem számítanának bele a képességbecslésbe, hiszen nem is feltétlenül megfelelőek, ugyanakkor nem lehet cél, hogy ez a mérés hosszát növelje, ez a paraméterek pontosságát ronthatja. Mivel az OKM-et évfolyamonként kb. 90 000 tanuló tölti ki, ezért egy-egy tesztbe elég lehet néhány próbaitemet elhelyezni. Egyetlen próbafeladat pozícióra véletlenszerű kiosztással legalább 84 000 / 300 = 260 itemet lehetne bemérni.

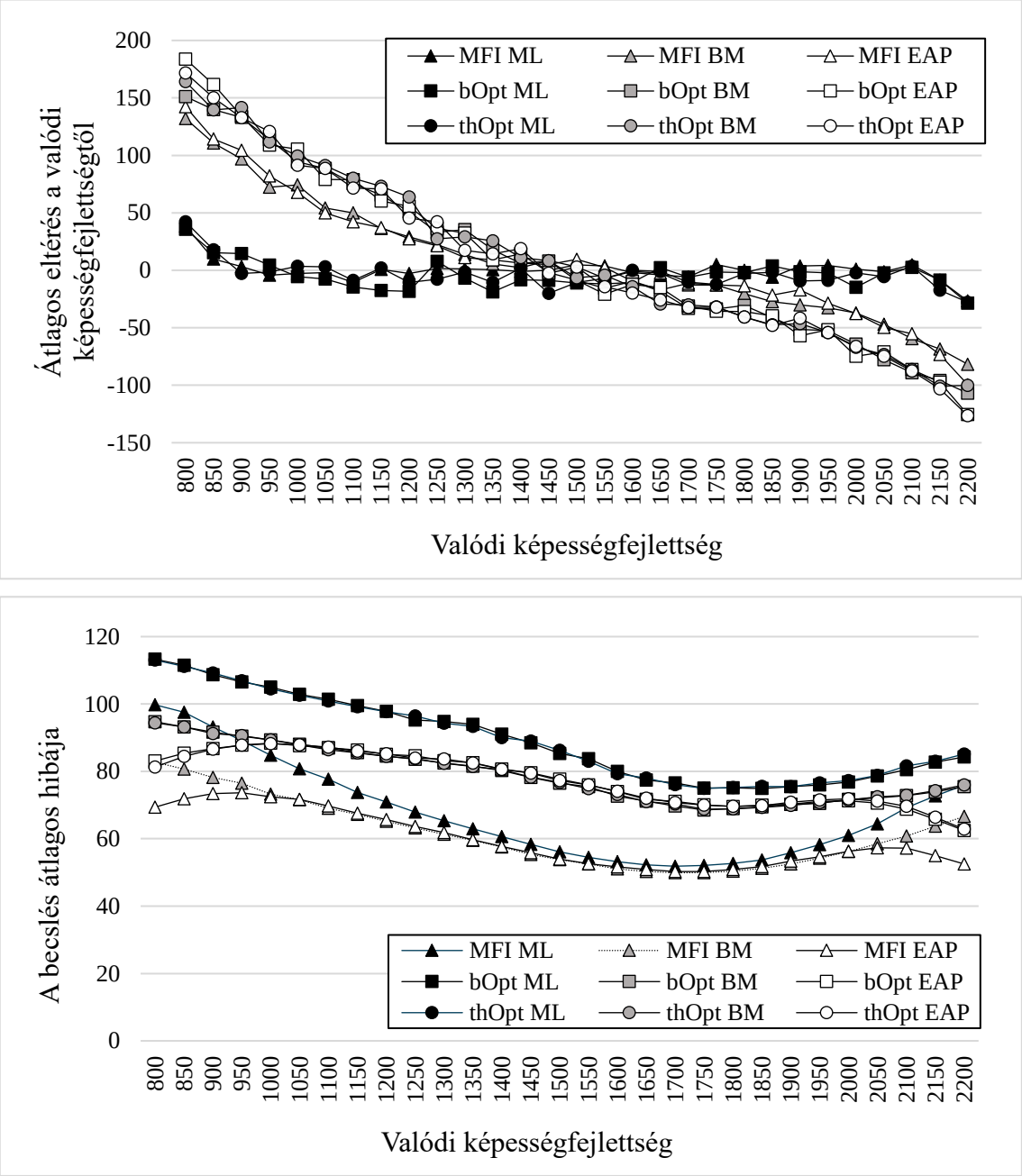
Első vizsgálatomban az adaptív mérés célja, azaz a teszt megállítási kritériuma a lineáris tesztnél valamivel rövidebb, 50 itemből álló teszt kitöltési folyamatának szimulációja. Ez a megállítási kritérium tesztfüzetenként akár 10 próbaitem vizsgálatát is megengedné, miközben a tanulók és az iskola terhelése nem változna a lineáris teszthez képest. Még öt próbaittemmel számolva is évfolyamonként évente 1300 új item bemérése válna potenciálisan lehetségessé, ez bőven fedezi a tesztfejlesztés szükségletét. A rövidebb teszthossz a próbaitemek felvétele nélkül a tanulói megterhelést egyéni szinten várhatóan csökkenti, azonban ekkor a próbamérés szervezés és tesztkitöltés szempontjából koncentráltabban jelenik meg a kiválasztott iskolákban.

Az egyes képességbecslési és item kiválasztási módszereket a képességfejlettség és a képességbecslés átlagos különbsége és a becslés hibájának átlagos nagysága alapján hasonlítottam össze.

Az eredmények alapján a képességfejlettség és a képességbecslés átlagos különbsége szerint a maximum likelihood becslések hozták a legjobb eredményt (15. ábra, felül). A képességskála szélein akár 100 ponttal kisebb eltérést mutattak, mint a Bayes-becslések. Ez nem meglepő abban az értelemben, hogy a Bayes-becslések minden esetben a priori eloszlás felé torzítanak, azaz esetünkben az 1500 pontos átlagérték felé. Az item kiválasztási módszerek közül a maximum Fisher információ szerinti kiválasztás valamivel jobb eredményt hozott, mint a legközelebbi nehézség és a legközelebbi maximális információ kiválasztási módszerek, azonban ez a különbség a képesség skála szélein nem tapasztalható (15. ábra, alul). Ezzel a módszerrel az 1200 és 1900 pont közötti tartományban jellemzően 40–60 pontnyi hibával terhelt tesztek születtek, ami körülbelül megfelel vagy valamivel jobb a jelenlegi lineáris teszt pontosságának. A másik két kiválasztási módszer esetén inkább 80 pontnyi hibával terhelt becslésekről beszélhetünk, de érdekes módon a legkisebb hiba nem 1500 pont környékén figyelhető meg, hanem ettől valamivel feljebb, 1600 és 1800 pont között.

15. ábra

A képességfejlettség és a becsült képességpont átlagos különbsége (felül) és a teszt várható hossza (alul) az egyes képességbecslési és itemkiválasztási módszerek szerint a képességskála finom felosztásán



Megjegyzés. Rögzített, 50 item hosszú teszt esetén.

MFI = Maximum Fisher információ, bOpt = legközelebbi nehézség, thOpt = legközelebbi maximális információ szerinti itemkiválasztás.

ML = maximum likelihood, BM = Bayes-modal, EAP = expected a-posteriori képességbecslési eljárás.

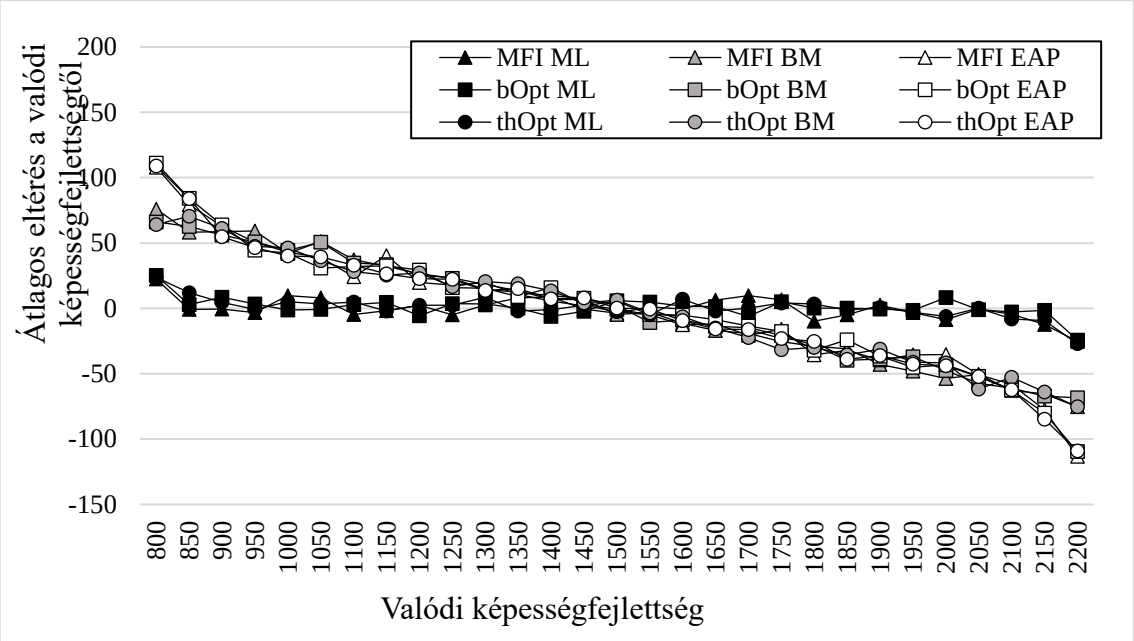
6.4.2. Rögzített hibahatárhoz tartozó várható teszthossz becslése

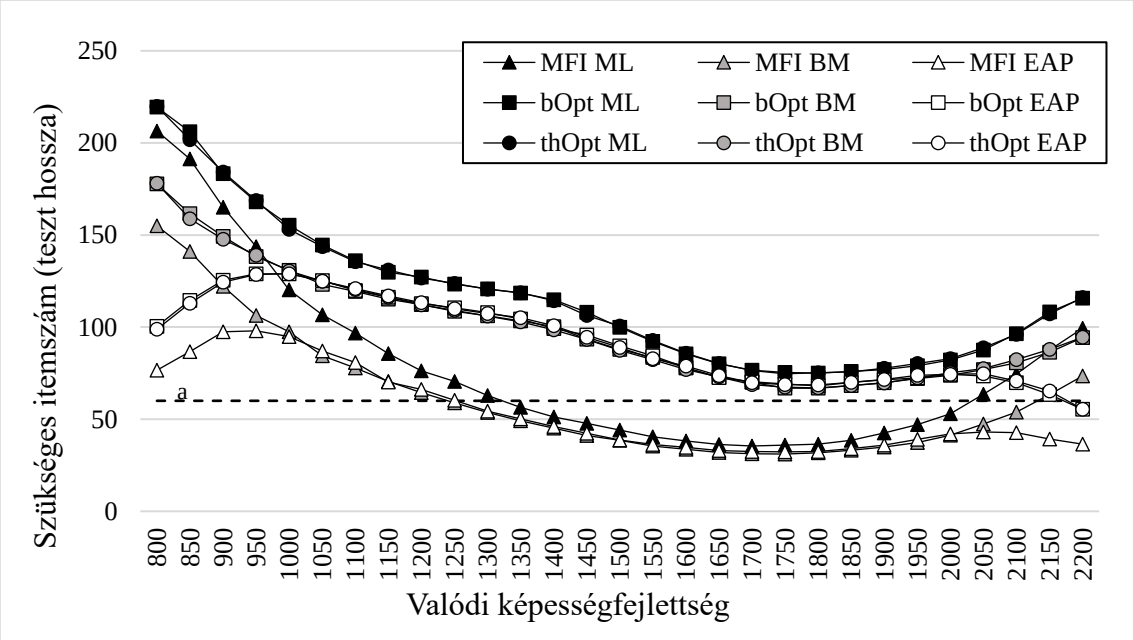
Második vizsgálatomban a teszt célját az előre meghatározott mérési pontosság elérése jelentette. A teszt addig folytatódott, amíg a képességbecslés becsült hibája el nem érte az előre meghatározott 60 pontot, a teljesítmény szórása 30 százalékát, ami megfelel az elvárt mérési pontosságnak, ha a tanulók egyéni szintjén értelmezhető eredmények elérése a cél (Kingsbury & Hauser, 2004). Az egyes képességbecslési és itemkiválasztási módszereket a képességfejlettség és a
képességbecslés átlagos különbsége és a tesztek átlagos hossza alapján hasonlítottam össze.

Az adott pontosságra törekvő tesztek esetében az előző szimulációhoz hasonló eredmények adódtak. A képességbecslési módszerek tekintetében ismét a maximum likelihood becslés hozta a legkisebb átlagos eltérést, ebben a vizsgálatban azonban a képességskála szélén sem volt nagyobb az elméleti képességfejlettség és a képességbecslés közötti különbség, mint 80 pont (16. ábra, felül). A tesztek hosszát tekintve szintén a maximum Fisher információ szerinti kiválasztás hozta a legkedvezőbb eredményt, ismét csak 1100 és 1900 pont között volt markáns eltérés a többi módszer eredményétől (16. ábra, alul). A lineáris tesznek megfelelő, 60 itemből álló teszthosszt 1300 és 2000 pont között sikerült rövidebb teszthosszra cserélni. Ebben a szimulációban is megfigyelhető volt, hogy nem a képességskála közepén, hanem attól valamivel feljebb kapjuk a legkedvezőbb eredményeket.

16. ábra

A képességfejlettség és a becsült képességpont átlagos különbsége (felül) és a teszt átlagos hossza (alul) az egyes képességbecslési és itemkiválasztási módszerek szerint a képességskála finom felosztásán





Megjegyzés. Meghatározott pontosságú, 60 pont egyéni becslési hiba esetén.

MFI = Maximum Fisher információ, bOpt = legközelebbi nehézség, thOpt = legközelebbi maximális információ szerinti itemkiválasztás.

ML = maximum likelihood, BM = Bayes-modal, EAP = expected a-posteriori képességbecslési eljárás.

6.4.3. Jobban diszkrimináló itembank esete

A harmadik vizsgálati szituációt az előző eredmények disszeminációja során felmerült ötlet ihlette[45]. Mint az előző fejezetben láthattuk, bár valamennyivel rövidíthető a teszt a papír-ceruza lineáris teszthez képest, és ez összhangban áll a 6.3 fejezet eredményeivel, ugyanakkor rögzített és elfogadható hibahatár mellett nem jelentős a teszt rövidülése. A Magyar Pszichológiai Társaság XXX. Országos Tudományos Nagygyűlésén elhangzott kérdés arra irányult, hogy lehetséges-e javítani a teszt hosszán jobban diszkrimináló, azaz meredekebb itemek segítségével. Ezek az itemek, habár a képességskála szűkebb tartományán szolgáltatnak információt, azonban jobban szétválasztják a nehézség alatti és feletti részébe tartozó kitöltőket. Ezért egy olyan vizsgálatot terveztem, melyben az itembank nehézség paramétereit továbbra is az OKM tartalmi keretéhez illesztettem, azonban a meredekséget az elfogadottnál nagyobbnak határoztam meg. Monte Carlo szimulációt alkalmaztam generált itembankkal és képességfejlettségekkel.

Vizsgálatom keretéül az OKM 2017. évi 6. évfolyamos matematikai eszköztudás mérésének item és tanulói szintű adatait használtam (Lak et al., 2018). Mivel mind a tanulói képességpontokat, mind az itemparamétereket az 1500 pont átlagú és 200 pont szórású skálán közlik, ezért az itemek és képességfejlettségek generálását is erre a skálára kalibráltam. A túl nehéz és túl könnyű feladatokat a tesztszerkesztési folyamat során eltávolítják (Auxné Bánfi et al., 2014), azonban a tanulói képességpontok esetében előfordulnak extrém értékek, ezért a képességpont-becslés tartományát a jelentésekhez igazodva 800 és 2200 pont közé korlátoztam.

Először egy lehetséges itembankot generáltam. Kétparaméteres modellt alkalmaztam, az itemek az OKM-nél szokásos nehézséggel, azonban a megszokottnál nagyobb diszkrimináló tulajdonsággal rendelkeztek. Az OKM tesztszerkesztése során a 0,3 és 1,7 logit közötti meredekségű itemeket fogadnak el (Auxné Bánfi et al., 2014). Vizsgálatomban fél ponttal emeltem a meredekséget, azaz a generált itemek meredeksége 0,8 és 2,2 logit közé esett. A paramétereket a transzformált skálához generáltam, így az itemek az OKM standard képességskáláján, 800 és 2200 között fognak mérni.

A mintanagyságot az OKM egy évfolyamához mérten 90000 főben határoztam meg, ezért az itembank nagysága 300 item (Magyar, 2014b). Ez az itembank-nagyság, az itemek 20%-os kitettsége esetén (azaz bármely itemet a kitöltők legfeljebb 20%-a lát) 60 itemből álló lineáris teszteket tesz lehetővé, azaz az adaptív teszt esetén legfeljebb 60

[45] Ezúton is köszönöm Nagybányai-Nagy Olivérnek felvetését.

hosszú tesztekkel a kritérium teljesíthető. A képességpontokat a fenti eloszlásból (N(1500, 200)) választottam.

Az itemparaméterek esetében a meredekségek minimum értéke 0,0019, a maximum értéke 0,0110 volt, az átlagos meredekség 0,0068. A nehézségparaméterek között a legkisebb érték 917, a legnagyobb 2190 pont, az átlagos nehézség 1537 volt. A generált képességfejlettségek minimuma 639,44, maximuma 2283,58, átlaga 1499,46 és szórása 200,56. Vizsgálatomban kétparaméteres modellt alkalmaztam, ezért a tippelési paraméter 0.

Itembank és elméleti képességpont birtokában számíthatók az egyes itemekhez tartozó megoldási valószínűségek és az item információs értékei. A megoldási valószínűség alapján adott képességfejlettséghez vagy képességfejlettségek vektorához generáltam minden itemhez egy-egy realizációt, tehát a függvény eredménye egy 90000 sorból és 300 oszlopból álló mátrix. Adott válaszmintázathoz elkészíthető a képességbecslés és a képességbecslés hibája. A válaszmintázatok generálása valódi véletlen abban az értelemben, hogy a képességbecslés akár jelentősen, egy szórásnyi mértékben eltérhet az elméleti képességponttól.

A kezdő fázis során egyetlen item kerül kiválasztásra (`nrItems = 1`), a belépési érték a populációs átlag (`theta = 1500`). Az első és további itemet maximum Fisher-információ elve alapján választjuk (`startSelect = "MFI"` és `itemSelect = "MFI"`) a legmegfelelőbb öt item közül (`randomesque = 5`). A teszt ciklikus része során Bayes-modal képesség becslést alkalmaztam (`method = "BM"`), ehhez a korábbi képességbecsléshez hasonlóan, beállítottam a szükséges prior és képességskála tulajdonságokat. A megállítási kritériumot úgy határoztam meg, hogy vagy az 50. item elérésekor vagy a képességbecslés hibája egy előre meghatározott értékének (60 képességpont, ami a szórás 30%-a) elérésekor érjen véget a teszt (`rule = c("length","precision"), thr = c(50, 60)`). A teszt végeztével az OKM-nél is használt expected aposteriori (`method = "EAP"`) becslést alkalmaztam, szintén az OKM-nek megfelelő beállításokkal.

Az itemek kitettségét 20%-ban határoztam meg, azaz egy itemet a kitöltők legfeljebb ötöde láthat. A szimulációkor egy post-hoc szimulációt futtatam, ahol a válaszmintázatok mátrixa a fentebb generált válaszrealizációkat tartalmazza, azonban a szimuláció egésze Monte Carlo szimulációnak számít, mivel ezek a válaszok szintén generálás eredményei, nem valódi adatfelvételből származnak. Ez a beállítás lehetővé

teszi, hogy pl. különböző képességbecslési vagy itemkiválasztási módszereket a válaszmintázatok véletlen hatásától függetlenül hasonlíthassunk össze.

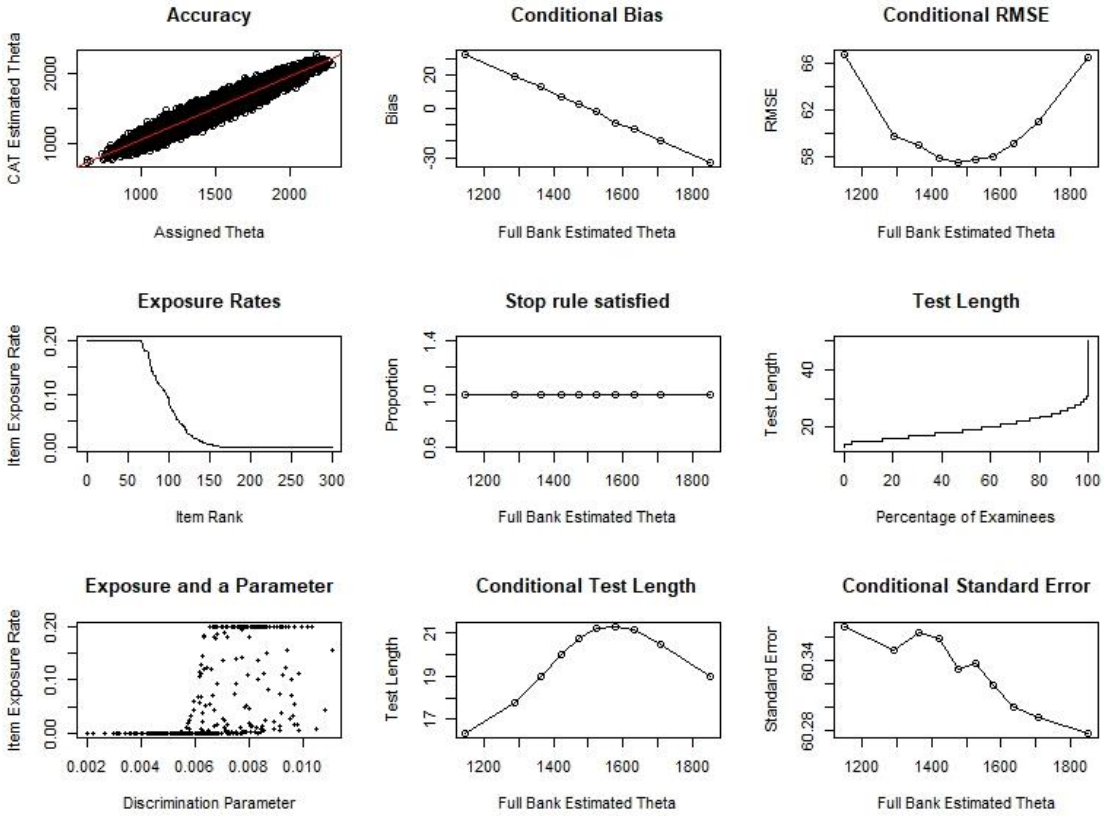
Az outputon (1. melléklet) alapján az elemzéssel eltelt idő közel két óra volt (Intel(R) Core(TM) i5-1155G7 @ 2.50GHz processzoron 8GB memória mellett). Az eredmények alapján a tesztek átlagosan 19–20 itemmel való foglalkozás után fejeződtek be. Az elméleti és a becsült képességpont közötti korreláció igen magas (r = 0,95). Az elméleti és a becsült érték közötti eltérés átlagos hibája 60 pont körüli (*RMSE* = 60,44), eltérések átlaga -0,39 pont. Az itemek kitettségét illetően, 44 item a lehető legnagyobb számban lett kiosztva, ugyanakkor 65 itemet egyáltalán nem használt föl a szimuláció. A tesztek átfedése 17,4%, ami a kitettségi arányok négyzetes összegének és a teszt hosszának hányadosa (Magis et al., 2017b, p.84).

A következő szekcióban az eredmények az elméleti képesség szerinti decilisekbe osztva találhatók. Az eredményeket a parancs három fájlba menti: az 1. megegyezik a konzolon megjelenített outputtal, a 2. a kilencvenezer teszt realizálóját tartalmazza, vagyis a kiválasztott itemek sorszámát, az így szimulált válaszmintázatot és az item utáni képességbecslés értékét. A .tables kiterjesztésű fájl minden sora egy-egy tanulói teszt összegzése, ami az elméleti képességfejlettség, a végső képességbecslés, a becslés hibája és a teszt hossza.

A futtatás eredményeiből többféle ábra is készül, amelyek egyesével is lekérhetők és menthetők. Az alapbeállítás ábráján (17. ábra) a felső sor három ábrája a teljes elemzésre vonatkozik. Az *Accuracy* az elméleti és becsült képesség pontok együtt járását, a *Conditional Bias* az elméleti és a becsült hiba különbségét, a *Conditional RMSE* pedig az átlagos négyzetes eltérést mutatja a teljes itembank alapján becsült képességpont szerint. Az *Exposure rates* és *Exposure and a parameter* ábrája az itemek kitettségét mutatja be. A felső arról tájékoztat, hogy az itemek milyen arányban lettek kiosztva, az alsó a kitettségi arányt a diszkriminációs paraméter függvényében ábrázolja. Ezek alapján megállapíthatjuk, hogy háromszáz itemnél jóval kevesebb is elegendő lenne a biztonságos működéshez, és elsősorban az alacsony megkülönböztető tulajdonsággal rendelkező itemek hagyhatók el. A *Stop rule satisfied* ábrája azt mutatja, hogy milyen arányban teljesülnek a megállítási kritériumok, de amikor a teszt hossza a megállítási kritérium, akkor jellemzően nem informatív.

17. ábra

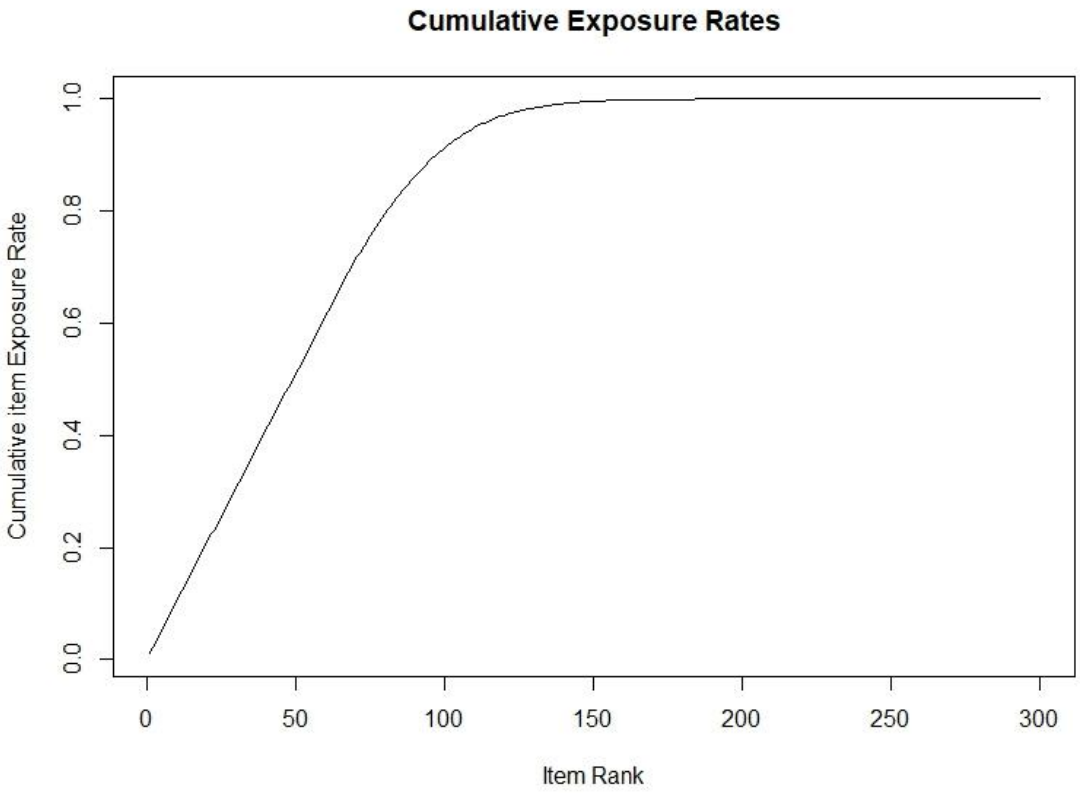
A teljes adaptív tesztet szimuláló `simulateRespondents` *függvény eredményének kilenc alapértelmezett ábrája*



A *Test lenght* ábrája azt mutatja, hogy a résztvevők jellemzően legfeljebb 30 item hosszú tesztekkel végeztek. A *Conditional Test Lenght* ábrája a teszt hosszát a teljes itembank alapján becsült képességponthoz viszonyítva mutatja, ami azt jelzi, hogy az 1500–1600 képességpont környékén voltak a leghosszabbak a tesztek. A *Conditional Standard Error* ábra a becsült képesség és a becslés hibájának kapcsolatát mutatja. A kitettség kumulált ábrája (18. ábra) szerint körülbelül 150 item, de legfeljebb 200 item elegendő volt a 90000 tesztkitöltő felméréséhez.

18. ábra

Az itemek kumulált kitettségének ábrája



Az eredmények alapján az OKM-ben megszokottnál meredekebb, vagyis jobban diszkrimináló itembank használatával lényegesen lerövidülhet a teszt hossza. A tesztek jelentős részében 30 itemnél kevesebb elegendő volt a teszt befejezéséhez. A teszt átlagos hossza 17 és 21 item között változott. Érdekes módon, a nagyon alacsony és nagyon magas képességfejlettség esetén rövidebb, az 1500 és 1600 közötti tartományban valamivel hosszabb tesztekre volt szükség, azonban ez a különbség is mindössze néhány item. Ez lényegesen rövidebb teszteket jelent, mint az előző szimulációban tapasztalt teszthossz, ami egyben azt is jelenti, hogy a képesség becslés hibájára vonatkozó megállítási feltétel teljesült, vagyis a tesztek jelentős részében a lineáris tesztnél lényegesen rövidebb adaptív teszt képes lehet elfogadható hibával terhelt képességbecslés elérésére.

Az adaptív teszt eredményeként kapott képességbecslések igen erősen együtt jártak az elméleti képességfejlettségek értékével. Az elméleti képességfejlettség és a becsült képesség közötti átlagos különbség a képességskála szélein körülbelül 30 pont volt, ami származhat a Bayes-módszerrel számított végső képességbecslésből is. Ez

egybevág korábbi szimulációs eredményeinkkel. A hiba nagysága a képességskála teljes egészén alig haladta meg a megállítási kritériumokban meghatározott 60 képességpontot. Érdekesség, hogy szimulációban a képességskála alsó részén valamivel magasabb, a felső részén pedig a valamivel alacsonyabb a hiba nagysága, azonban ez a különbség elenyésző, egy ponton belüli.

A meredekség paraméter szerinti kiválasztás aránya a magasan diszkrimináló itemek értékében jellemzően eléri a maximális 20%-ot, míg a kevésbé jól diszkrimináló itemek esetében számos item egyáltalán nem került mérésbe. Az itemek kumulált kiválasztási ábrája alapján ekkora tanulói bázisnál elegendő lehet a háromszáz itemnél kisebb, 150–200 itemből álló itembank, feltéve, hogy az itemek diszkrimináló képessége magas.

A nagyobb meredekséggel rendelkező itemek kiválasztására a valószínű magyarázat az MFI kiválasztási módszerben rejlik, mely adott képességbecslés esetén a legnagyobb információval rendelkező itemet választja. Amennyiben a képességbecslés közel esik az item nehézségéhez, a meredekebb itemnek nagyobb lesz az információja, jobban szétválasztja a nehézség-környéki képességbecsléssel rendelkező kitöltőket. A szimuláció során véletlenítést alkalmaztunk (1. melléklet), azaz adott képességbecslés esetén az öt legnagyobb információval rendelkező item közül véletlenszerűen kerül kiválasztásra a következő feladat, így a közeli, nagy diszkriminálóképességű itemek mellett kisebb meredekségű feladatok is kiválasztásra kerülhettek, míg meredekebb, de nehézségükben a képességbecsléstől távolabbi itemek nem kerültek kiválasztásra.

7. Összegzés

Kutatásomban az Országos kompetenciamérés fejlesztésének digitalizálásban rejlő egyik irány, a számítógépes adaptív teszt módszertani kérdéseit vizsgálatam. Azaz olyan megelőző vizsgálatokat, melyek előkészítik az egyik hazai tanulóiteljesítmény-mérési rendszer esetében a szakmai és mérésmódszertani szempontból sikeres papír-ceruza – számítógépes adaptív adatfelvétel átmenetet. A papír-ceruza teszt és a számítógépes mérés közötti átmenet a kutatás ideje alatt megvalósult[46], azonban a nemzetközi nagymintás tanulóiteljesítmény-mérések médiahatással kapcsolatos eredményeink összegzése továbbra is hiánypótlónak számít. Az OKM elkötelezett az adaptív mérési módszer fejlesztése mellett mind szakmai (Balázsi et al., 2021), mind oktatáspolitikai (Karkó, 2023) oldalról, ezért a lineáris tesztről az adaptív mérésre történő áttérés vizsgálata aktuális és releváns, az OKM-ből származó információk széleskörű használata miatt a téma vizsgálata társadalmi hasznossággal bír.

7.1. Eredmények a kutatási kérdések tükrében

Kutatásomban 3 fő kérdésre és azok alkérdéseire kerestem a választ. A továbbiakban az ezekre kapott eredményeket összegzem.

1) Az OKM papír-ceruza méréseiből származó adatok relevánsan felhasználhatók-e a számítógépes adaptív mérés tervezésére?

a. *Mi a számítógépes mérési környezet hatása a mérés eredményére? Kell-e médiahatásra számítani, és ha igen, hogyan kezelhető? (6.1 fejezet)*

A médiahatás-vizsgálat a PISA mérés esetében a 2015. évi mérés próbamérésének keretében (OECD, 2016a, 2017b), a TIMSS 2019 esetében pedig önálló vizsgálattal történt (Fishbein et al., 2018). A vizsgálatok a teljes mérésre koncentrálnak, azon belül céljuk 1) a két mérés konstruktum-azonosságának vizsgálata, 2) az itemek szintjén történő médiahatás-vizsgálat, és 3) a tanulói teljesítmények szintjén történő médiahatás-vizsgálat, egyszersmind a trendek folytonosságának biztosítása áll. Mindkét vizsgálat megerősíti, hogy a két tesztelési módban mért konstruktum azonosnak tekinthető, csak az itemek kis részében adódik szignifikáns különbség az itemparaméterekben. Míg a PISA

[46] Az OKM 2021-ig papír-ceruza teszt, 2022-től számítógépes mérés formájában került megszervezésre (Oktatási Hivatal, 2023a).

mérés a két módban nem invariáns itemek esetében mindkét irányú médiahatásról beszámol, addig a TIMSS vizsgálat jellemzően a számítógépes adatfelvételt találta nehezebbnek. Az itemek meredekségében nem mutattak ki médiahatást, azaz az itemek viselkedése azonos a két módban. Mindkét mérés esetében szükségesnek találták valamilyen korrekció alkalmazását, azonban mindössze néhány itemnél határoztak meg országspecifikus item paramétereket, amely azonban nem volt szükséges minden területen, azaz a médiahatás nem országfüggő. A teljesítmények összességében 10–20 pontnyi eltérést mutatnak az egyes területeken (és évfolyamokon), ezért a trendek számításában érvényesítik a médiahatás korrekcióját. Ekkora különbség körülbelül 6–8 helyezésnek felel meg a mérések rangsoraiban (a középértékek közelében).

A nemzetközi adatbázisokban végzett szisztematikus keresés során nyolc publikációt találtam, melyek a médiahatást valamely mérés adatbázisai vagy feladatai segítségével vizsgálják. Ezek jellemzően abból a szempontból közelítik a kérdést, hogy az egyes országok esetében eltérő médiahatással és ennek megfelelő trend-korrekcióval kell-e számolni. Az eredmények, akárcsak a médiahatás vizsgálatának általános eredményei, eltérő képet mutatnak. Úgy tűnik, hogy a médiahatás bizonyos országoknál negatív, másoknál pozitív irányú lehet a papír-ceruza mérés eredményeihez képest (Jerrim, 2016), de az is előfordul, hogy nincs kimutatható médiahatás (Hamhuis et al., 2020). Ennek következménye lehet, hogy egyes országok trendje alá- vagy felülbecsli a valóságos eredményt, bár ez a torzítás jellemzően nem jelentős. A médiahatás kimutatható a számítógépes adatfelvétel szöveges válaszainak nagyobb elem-gazdagságában és információtartalmában is (Zehner et al., 2019, 2020).

b. *A nyílt végű itemek elhagyása mellett is azonos marad-e az OKM mérés tartalmi kerete? Vagyis, ha kizárólag zárt végű itemeket alkalmaznánk, akkor milyen eltéréseket tapasztalnánk a tanulók képességpontjának becslésében? (6.2 fejezet)*

Vizsgálatomban a teljes, azaz a nyílt és zárt itemeket is tartalmazó tesztből származó teljesítményt vetettem össze egy olyan, rövidebb tesztből származó teljesítménnyel, amelyet csak a zárt végű itemekből számítottam. A vizsgálatot az OKM 2017. évi méréséből származó adatokon, 6., 8. és 10. évfolyamon a szövegértés és matematikai eszköztudás területeken végeztem. Számításaim igazolták, hogy meglehetősen közeli, 0,9 feletti korrelációs szintű egyezést lehet kimutatni a teljes teszt és a csak zárt itemek alapján számított képességpontok között. Ez arra utal, hogy a nyílt

itemekkel mért konstruktum megegyezik a teljes teszt által mért konstruktummal, a nyílt itemek nem valamely más kérdésformával nem mérhető gondolkodási műveletet vagy tartalmi területet mérnek.

A folytonos teljesítményt az OKM módszertanának megfelelően képességszintekbe osztva, majd a két teljesítményből számított képességszintet összevetve kiderült, hogy a csak zárt kérdések segítségével készült szintekre való besorolás 8 képességszint esetén 2 szintnél nagyobb arányú tévesztést az esetek kevesebb, mint 1 százalékában eredményez. A kategória szintű vizsgálatok azt mutatták, hogy jellemzően a mérési skála két végletén tapasztalhatók jelentősebb eltérések, ami egybevág Geer (1991) eredményeivel. A magasabb képességszinteken teljesítők esetében lefelé, a legalsó képességszinteken teljesítők esetében felfelé torzítás észlelhető a nyílt végű itemek (képzett kódoló munkáját igénylő feladatok) elhagyásával. Ez alapján a szöveges választ kívánó feladatok elhagyása a felsőbb tartományban a valóban jó képességgel rendelkezőket némileg alulértékeli, ugyanakkor a teljesítmény skála alsó régióiban éppen ezzel ellentétes torzításokkal járt együtt, a rosszabb képességű kitöltők hiányosságai rejtve maradhatnak.

Első két alkérdésem arra irányult, hogy a papír-ceruza teszt számítógépes tesztre történő cseréjének lehet-e olyan következménye, ami miatt az OKM eddigi méréseinek item és tanulói szintű eredményeit nem lehet adaptív számítógépes mérés tervezésére felhasználni. Az eredmények szerint magának a médiumnak a cseréje más, hasonló célú és módszertanú méréseknél 1) nem okozott konstruktum-validitási problémát, 2) az itemek jellemzői nagy hasonlóságot mutatnak a papír-ceruza itemek jellemzőivel, 3) a teljesítményben nincs jelentős különbség, és az OKM-en végzett vizsgálat alapján 4) a nyílt itemek elhagyása, azaz csak zárt itemek használata hasonló eredményre vezet, mint a teljes teszt.

Ha elfogadjuk, hogy a nyílt végű feladatok egy része a papír-ceruza formához hasonló zárt formára alakítható, ahogy az OKM 2022. évi felmérésében is történt, akkor szimulációinkban a nyílt itemeket és azok jellemzőit szintén felhasználhatjuk.

2) Az eredeti méréssel megegyező mérési pontosság mellett mi az adaptív teszteléssel elérhető legrövidebb teszthossz? (6.3 fejezet)

Második kutatási kérdésem arra irányult, hogy az adaptív mérés szimulációs tervezési fázisához viszonyítási pontokat határozzak meg. Az elméleti levezetés célja

olyan képlet megalkotása volt, melynek segítségével előre meghatározott tulajdonságokkal – úgymint mintanagyság, item mérési pontossága, teljesítmény mérési pontossága, mérés finomsága – rendelkező méréshez meghatározzam a 1) próbamérés szükséges minimális mintanagyságát és 2) az adaptív mérés átlagos minimális hosszát. A levezetés néhány egyszerűsítéssel élt, mivel az IRT modellek közül a Rasch-modellt alkalmaztam, illetve a mérés finomságát a pontos itemnehézségek és teljesítmény helyett képeségszintekkel mértem. Ez a megközelítés nem áll nagyon távol a CAT lényegétől, megfelel a klasszifikációs célú méréseknek, ahol a mérés célja a teszt kitöltőjének elhelyezése *K* meghatározott tudásszint valamelyikén. Hasonló célja az OKM-nek és a nemzetközi nagymintás tanulóiteljesítmény-méréseknek is van, amikor a tanulókat a teljesítmény alapján képességszintekbe osztják. Az adaptív teszt során a kitöltő valamely képességszintre van besorolva, és a következő itemet a neki megfelelő nehézségű itemszintről kapja. Az optimális próbaméréshez feltételeztem továbbá, hogy a kitöltők képességszintjét a mérés eredményeként ismerjük, így bármely bemérésre szánt itemet kioszthatunk meghatározott számú és meghatározott képességszintű kitöltőnek. Ezt több képességszinttel is megtehetjük, ezzel természetesen a próbamérés nagyságát növelve.

Az OKM mérési keretének megfelelő feltételek szerint a mintanagyságot 100 000 főben, a képességszintek számát 7 szintben, az itemek mérési hibáját a teljesítmény szórásának 20 százalékában határoztam meg, a teljesítmény hibáját pedig a teljesítmény szórásának 50-60 százalékában korlátoztam. Emellett a teszt tervezetten átlagos nehézségű volt, vagyis az itemeket úgy osztottam ki, hogy a megoldás valószínűsége lehetőleg 50% legyen. Ekkor 1) az item bemérésekor minden olyan képességszinten, amelyen az item viselkedését bemérjük, 377 főt igényel (17. táblázat), és 2) az egyes kitöltők esetében az adaptív teszt bármely képességszinten átlagosan 41–58 item hosszú (18. táblázat).

Ez az eredmény a korlátok ellenére újszerű, eddig hiányzó hidat képez a modellek elméleti egyenletei és a gyakorlat között. A képlet alapján megfelelő itembankkal az OKM esetében minden képességszinten, még a legalacsonyabb és legmagasabb szinteken is, teljesíthető a megfelelő pontosságú teljesítménymérés legalább a lineáris tesztnek megfelelő teszthosszal. A képességszintek számának csökkentésével, kisebb pontosságú klasszifikációval ezek a mérőszámok drasztikusan csökkenthetők, pl. 5 képességszint esetén 44 item elegendő. A pontosságot a teljesítménybecslés irányából csökkentve nagyobb, 60 százalékos becslési hiba mellett már 41 item hosszú tesztekre számíthatunk.

Az elemzés azonban nem használja ki a többparaméteres modellekkel, a meredekebb itemek használatával járó előnyöket. A kialakított egyenletek ugyanakkor lehetőséget biztosítanak arra, hogy az átlagostól eltérő nehézségű, a tesztkitöltés motivációját jobban segítő könnyebb tesztek tervezését segítse. Ekkor a próbamérés során nagyobb mintanagysággal és hosszabb tesztekkel kell számolni.

3) Az OKM papír-ceruza méréseiből származó adatok alapján mely adaptív mérési elemek valószínűsítik a mérés céljának sikeresebb megvalósítását (a matematikai eszköztudás területen)? (6.4 fejezet)

a. *A papíralapú teszttel azonos itemszám mellett az adaptív teszt esetében csökken-e a tanulói képességfejlettség-becslés hibája? (6.4.1 fejezet)*
b. *Az OKM adatain alapuló, számítógépes adaptív tesztet imitáló szimulációk alátámasztják-e, hogy lényegesen rövidebb idő alatt (kevesebb itemmel) a papír-ceruza teszt pontosságának megfelelő pontossággal meghatározható a diákok képességpontja/teljesítménye? (6.4.2 fejezet)*
c. *Milyen megállítási kritériumok milyen mérési céloknak felelnek meg az adaptív OKM kapcsán?*
d. *A megállítási kritériumok között van-e hierarchia, azaz léteznek-e olyan erős kritériumok, melyek teljesülése magával hozza a gyengébb feltételek teljesülését?*
e. *Az első 5–10–15–20 kérdés után változik-e még a diák teljesítménye? 5–10 kérdéses teszthossz mellett milyen teljesítménybecslések várhatók?*

Harmadik kutatási kérdésem az OKM adaptív megvalósítására, az egyes tesztelemek alkalmazhatóságára irányult. Ezek a vizsgálatok Thompson és Weiss (2011) keretrendszere szerint az első lépéséi a számítógépes adaptív tesztek fejlesztésének, és ehhez a szakaszhoz a leginkább a Monte Carlo szimulációk illeszkednek. A szimulációk keretét a 6. évfolyamos matematikai eszköztudás mérése adta. A belépési értéket a 2010. évi populációs átlagnak megfelelő 1500 pontos értékben határoztam meg. Ez a legelemibb lehetőség a belépési érték meghatározására, annak a mérési szituációnak felel meg, amikor a teszt kitöltőjéről nincs olyan információ (demográfiai jellemző, korábbi mérés eredménye vagy más területen elért teljesítmény), ami segítené a teljesítménye előzetes becslését.

A megállítási kritériumok vizsgálatára irányuló *c.* alkérdés keretében két valószínű mérési célt tűztem ki. Az első változat szerint minden tanuló rögzített

hosszúságú, 50 itemből álló tesztet tölt ki, ami valamivel rövidebb, mint az eddigi mérések lineáris tesztjei[47], ezért a fennmaradó időt az ismeretlen paraméterekkel rendelkező próbaitemek tehetik ki. A *2)* kutatási kérdésre kapott válaszok alapján ez a teszt várhatóan elegendően pontos képességbecslést eredményez, valamint lehetőséget adna arra, hogy az új itemeket minden szükséges képességszinten azonos nagyságú minta alapján értékeljük.

A második mérési cél esetében a teszt akkor ér véget, amikor a kitöltő képességbecslésének hibája egy előre meghatározott határ, 60–70 pont alá csökken. A határértéket a lineáris tesztek teljesítményeire számított hiba alapján választottam ki. A képességskála közepén, jellemzően ennél alacsonyabb (30–40 pont), a szélein ennél lényegesen magasabb (100–120 pont) hibák a jellemzőek. A 60–70 pontnyi hiba alacsonynak mondható, a teljesítmény szórásának (200 pont) mindössze 30–33 százaléka, vagyis alacsonyabb, mint a *2)* kutatási kérdésben vizsgált 40–60 százalék. A második megállítási kritérium tehát elsősorban a nagy pontosságú mérést célozza meg, kevesebb teret engedve a főméréshez kapcsolódó próbamérésnek.

Az *a.* és *b.* alkérdésekre irányuló szimulációk eredménye alapján a két mérési cél hasonló eredménnyel zárulna. Mindkét esetben azt találtuk, hogy az 1300 és 2000 képességpont esetében a két megállítási kritérium megfeleltethető egymásnak, azaz a képességpontok környékén a 60 pontnyi hiba 50 hosszú tesztekkel érhető el. Az 1300–2000 képességpont közötti tartományban a tesztek jobban teljesítenek a másik megállítási kritériumnál, vagyis az 50 hosszú tesztek 60 pontnál kisebb hibával zárulnak, a 60 pontnyi hiba eléréséhez elegendő 50-nél kevesebb item. A képességskála szélén kissé más a helyzet: 50 item hosszú tesztekkel még a legalacsonyabb képességfejlettség esetén sem kapunk 70–90 pontnál nagyobb becslési hibát, ami a teljesítmény szórásának 35–45 százaléka, és megfelel a lineáris teszten kapott hiba nagyságának. A 60 pontos hibahatár eléréséhez a legjobb beállításokkal is 75–100 itemre lenne szükség, ami a jelenlegi lineáris teszteknél lényegesen hosszabb. A képességskála felső szélén hasonló a helyzet: a jelenlegi itembankra építve 50 itemmel a lineáris teszt eredménye elérhető, azonban a 60 pontos hibahatár csak a lineáris tesztnél hosszabb adaptív teszttel megvalósítható. Ez alapján a *d.* alkérdésre a válasz, hogy a megállítási kritériumok kevert stratégiája hozhatja a legjobb eredményt: a teszt akkor ér véget, amikor eléri a teljesítmény hibája a 60 pontos határt vagy a teszt 50 item hosszú. Ez a kombináció kellően pontos teljesítménybecslést

[47] 2008 és 2019 között a 6. évfolyamos matematika tesztek 55 és 60 item közötti hosszúságúak voltak.

ad egy lehetséges próbamérés kiosztásához, valamint a képességskála középső részén, ahol a nagyobb létszám és a belépési érték megválasztása miatt több és jó minőségű itemre van szükség, a több fennmaradó item miatt több próbaitem kipróbálására van lehetőség.

A számítógépes tesztelés hat eleme (IRT modell, itembank, belépési érték, itemkiválasztás módja, teljesítménybecslési eljárás, megállítási kritériumok) közül szimulációim elsősorban a teljesítménybecslési eljárások és az itemkiválasztási módszerek összehasonlítására irányult. A vizsgálat során maximum likelihood és Bayes típusú teljesítménybecsléseket, valamint maximális információn és nehézségparaméteren alapuló itemkiválasztási módszereket hasonlítottam össze. A képességskála egyenletes felosztásán szimuláltam adaptív teszteket, az egyes módszereket a válaszok generálásához használt képességpont és a képességbecslés különbségével és a teljesítményhez számított hiba nagysága vagy a teszt várható hossza alapján hasonlítottam össze.

Mindkét megállítási kritérium esetében hasonló eredményt kaptam. Az elméleti képességponttól való távolság alapján a maximum likelihood becslés bizonyult pontosabbnak. A Bayes-becsléseket elsősorban nagyon rövid tesztek vagy adaptív tesztek elején célszerű alkalmazni, ahol a centráló tulajdonság stabilizálja a képességbecslést. Hosszabb tesztek esetén a középre torzítás inkább hátránynak tűnik. Az eredmények alapján az 50 itemből álló vagy a 60 pontnyi hibát elérő teszt elegendően hosszú ahhoz, hogy a maximum likelihood becslés kedvezőbb legyen. Az itemkiválasztási módszereknek nincs jelentősége a képességpont és a képességbecslés különbségének szempontjából.

Az itemkiválasztási módszerek közül mindkét szimulációban a maximum Fisher-információ szerinti kiválasztás bizonyult jobbnak. Ez a módszer jobban kihasználja a többparaméteres IRT modellekben rejlő lehetőségeket, például a nagyobb meredekségű itemek jobban diszkrimináló tulajdonságait. Ugyanakkor a maximum Fisher-információn alapuló módszernek nagyobb tárigénye vagy számítási igénye van, mint az itemenként egyetlen jellemző vizsgálatát igénylő legközelebbi nehézség vagy legközelebbi maximális információ alapú kiválasztásnak. A teljesítménybecslési módszerek között nem találtam jelentős különbséget a teljesítménybecslés hibájának szempontjából, bár az expected a-posteriori becslés a képességskála legszélén valamivel kisebb hibát generált.

Eredményeim alapján a *3)* kutatási kérdésre adott válasz, hogy az OKM adaptív megvalósításában a maximum likelihood teljesítménybecslési módszer és a maximum

Fisher-információ szerinti itemkiválasztási módszer használatát javaslom, mivel a teljes képességskálát egyforma fontossággal kezelő eredmények alapján ennek a két módszernek a kombinációja biztosítja a legkisebb eltérést a képességfejlettség és a képességbecslés között, valamint a teljesítménybecslés legkisebb hibáját, vagyis összességében a legpontosabb teljesítménybecslést.

Szimulációmban feltételeztem, hogy az adaptív OKM itembankja felhasználná a korábbi mérések beszerkesztett és bemért itemeit[48]. Erre a feltevésre alapozva az *a.* és *b.* alkérdés vizsgálatára irányuló szimulációk itembankját a 2008 és 2019 közötti mérések 625 itemének paraméterei alkották. A teljesítmény hibájára vonatkozó eredmények azonban némi egyenetlenségről tanúskodnak, a középső, 1500 képességpont körüli tartomány helyett a magasabb, 1500–1900 közötti tartományban a legkisebb a becslés hibája vagy a szükséges itemek száma[49]. Ennek több lehetséges magyarázata lehet. A 6. évfolyamos mérésben használt itemek egy része magasabb évfolyamokon is szerepel, így elképzelhető, hogy a képességskála ezen részén gazdagabb az itembank. Másrészt az OKM itemei a mérés céljának megfelelően a képességskála középső részére, a nagy számú átlagos képességfejlettségű tanuló mérésére koncentrál, ezért a valódi itemek esetében kevés nagyon alacsony vagy nagyon magas nehézségű item van, ami rontja a megfelelő item kiválasztását az alacsony és magas képességfejlettség esetén, tehát nagyobb hibát eredményez. Harmadrészt, az OKM itemeit úgy válogatják, hogy a képességskála minél nagyobb részén, vagyis lehetőleg több tanuló esetében szolgáljanak információval, ezért az itemek meredeksége inkább alacsonynak mondható. Az adaptív mérés, és különösen a maximum Fisher-információ itemkiválasztási módszer esetén azonban a magasabb meredekségű, jobban diszkrimináló itemek használata előnyösebb.

Utolsó szimulációnkban az OKM jelenlegi itembankját egy 300 itemből álló, a képességskálát nehézség szerint egyenletesen fedő, de a meredekség szempontjából meredekebb itemeket tartalmazó feladatbankra cseréltem. A kiválasztás a maximum Fisher-információ módszerével történt, a megállítási kritérium 50 item vagy 60 pontnyi hiba elérése volt. A képességfejlettség és a becslés különbsége, valamint a becslés hibája a korábbi szimulációkhoz hasonló nagyságrendű volt, azonban a teszt hossza jelentősen lerövidült, az esetek 60 százalékában legfeljebb 20 item volt szükséges. A korábbi szimulációktól eltérő eredmény, hogy a leghosszabb tesztek nem a képességskála szélén,

[48] A 2022-es számítógépes mérés során így is történt, a papír-ceruza és a számítógépes mérés közötti összekötést korábban megbízhatónak bizonyult itemek biztosítják (Balázsi et al., 2021).
[49] A képességpont és a képességbecslés különbsége esetén nem találtam ilyen eltérést.

hanem az 1500–1700 közötti tartományban adódtak. Ez alapján valószínűsíthető, hogy az OKM jelenlegi itembankjában több alacsony és magas nehézségű itemre van szükség. További eredmény, hogy a feladatbankból valóban a legnagyobb meredekségű itemek kerültek leggyakrabban kiosztásra, méghozzá egy évfolyamnyi (90 000) kitöltő esetén mindössze 150 itemre volt érdemben szükség. Ez alapján a *3)* kutatási kérdésre adott válasz itembankra vonatkozó része, hogy az adaptív OKM itembankjának fejlesztésekor a magasabban diszkrimináló, korábban esetleg ebből a szempontból alkalmatlannak ítélt feladatok jól megválogatott szettje lehet sikeres.

7.2. A kutatás korlátai és kitekintés

A médiahatás vizsgálat érvényességének legnagyobb korlátja, hogy nem az OKM adatain végzett empirikus vizsgálat, hanem a módszertanukban példaképként szolgáló nemzetközi nagymintás tanulóiteljesítmény-mérések eredményei. Ezek a vizsgálatok, a különböző oktatási rendszerek mérésére és összehasonlítására koncentrálva, elsősorban arra koncentrálnak, hogy az egyes országokban jelentősen más médiahatással kell-e számolni. Az OKM esetében erről nincs szó, azonban a magyarországi tanulók esetében is indokolt lehet alcsoportok vizsgálata, amihez a nemzetközi eljárások példaként szolgálhatnak. További kutatási korlát lehet, hogy kizárólag a PISA, TIMSS és PIRLS mérésekre koncentráltunk, így pl. az USA mérési rendszereinek tapasztalatai nem kerültek feldolgozásra. Általában véve a most feltárt eredmények elsősorban nyugat-európai megközelítésből tárgyalják a kérdést, így egy későbbi körben érdemes lehet amerikai vagy távol-keleti és ausztrál, valamint kifejezetten európai országos mérések tapasztalataira koncentrálni.

A nyílt itemek elhagyása következményeiről szóló eredmény kizárólag a 2017. évi mérés eredményein alapul. Lehetséges lenne több mérési év vizsgálata is, bár jelentősen eltérő eredményre nem számítok. Fontosabb korlátnak tekintem, hogy nem két empirikus adatfelvétel, egy teljes és csak zárt itemeket tartalmazó teszt eredményének összehasonlítása, hanem teljes teszt és ugyanazon teszt rövidített változatából származó eredmények összevetése történt. A vizsgálat nem számol a nyílt itemek teszten belüli elhelyezkedésével vagy nehézségével, vagy a rövidebb teszt során várható kisebb fáradással. Szintén nem mérlegeltem a nyílt végű feladatokat abból a szempontból, hogy mennyire könnyen, akár változtatás nélkül alakíthatók-e automatikus kódolású zárt itemekké.

Az OKM 2022. évi mérésében szerepeltek automatikus kódolású nyílt itemek és képzett kódoló munkáját igénylő nyílt itemek is. A 2023. évi mérésben már kizárólag automatikus kódolású itemek voltak, így mostanra feltételezhetően kialakult a digitális itemek szerkesztésének módszertana. Ennek fényében érdemes lenne a vizsgálatot empirikus adatfelvétellel megismételni, akár többféle tesztverzióval, ahol némelyikben kódolandó nyílt itemek is szerepelnek.

Az adaptív tesztek elméleti optimumával kapcsolatos eredmény, bár hidat képez az elméleti egyenletek és a gyakorlat, például a szimulációk között, korlátozott érvényességű. Először is, az egyenletek a Rasch-modellen alapulnak, holott a mai tanulóiteljesítmény-mérések jellemzően többparaméteres modelleket használnak. Ugyanakkor a többparaméteres egyenletekből származó levezetések lényegesen bonyolultabbak, további kutatást igényelnek, ha egyáltalán megoldható feladatra vezetnek. Második, gyakorlatibb jellegű korlát, hogy a levezetésben idealisztikus körülményeket feltételeztem, például hogy mindig van pontosan a képességbecslésnek megfelelő item, vagy minden lépésben pontosan illeszkedik az item megoldási valószínűségének valódi és becsült értéke. Ez a feltétel a teszt elején nagy valószínűséggel nem teljesül. Harmadik, strukturális jellegű megjegyzésem, hogy a levezetésben a CAT és az MST egyfajta keverékét modelleztem, vagyis a mérés finomságát az MST-hez hasonló szintekkel szabályoztam, míg a teszt hosszát a CAT egyik megállítási kritériumához hasonlóan a becslés pontossága határozta meg. Továbblépési lehetőség lenne, hogy a többszakaszos tesztek esetében a szükséges szakaszok vagy az egyes szakaszokban helyet kapó modulok hosszára alkossunk hasonló elméleti modellt, illetve a számítógépes adaptív tesztelés esetén a szintekkel szabályozott mérési finomságot folytonos változatra cseréljük.

Az OKM-ből származó valódi adatokat felhasználó eddigi szimulációs vizsgálatok (6.4.1 és 6.4.2 fejezetek) csak az itemek paramétereit használták fel, és a képességskála teljes terjedelmét ugyanolyan súllyal vették figyelembe. Emiatt a Bayes-módszerek eredményessége kisebbnek tűnhet. Kizárólag az itemparaméterek használata azt is jelenti, hogy a tanulói adatfájlok (teljesítménypontok és válaszpontszámok) használata elmaradt. Továbblépésként tervezem, hogy a valódi tanulói teljesítménypontokkal, mint valódi teljesítménnyel, teljes teszt szimulációját végezzem. Egy ilyen vizsgálat jobban közelítené a teljesítménybecslések valódi mérésen történő teljesítését. A valódi teljesítmények használata esetén azonban a tesztfüzetek leválogatásának következményeivel is számolni kell. A tanulói adatfájlok esetében a

hiányzó és üres füzetek, illetve a mentesülő és a tesztet idegen nyelven író tanulók füzetei nem kerülnek elemzésre. Első esetben erre nincs is lehetőség adatok hiányában, utóbbi esetben éppen az országos eredmények torzításának elkerülése a cél. Kutatásom szintén kizárja ezeket a tesztfüzeteket, ez konzisztensé teszi mérése tárgyával. A hiányzó és üres füzetek az eredeti mérés során is vezethetnek torzításhoz, amennyiben nem véletlenszerűen, hanem szisztematikusan jelennek meg a populációban. Amennyiben ez a helyzet, úgy a számítógépes és adaptív mérés esetében is ez történhet, így nem okoz validitási problémát. Valójában az új típusú mérés esetén, ha egyetlen mérési nap helyett mérési időszakban valósul meg, a hiányzó tanulók számának is csökkennie kellene.

A kutatásban többletként alkalmazott megszorítás (csak mindkét tesztrészen jelenlevő tanulók bevonása) szintén a minta csökkenéséhez vezet, évfolyamonként legfeljebb néhány száz tesztfüzet kizárásához, ami elhanyagolható a minta méretéhez, az évfolyamonként hatvanezer tanulói sorhoz képest, és véletlen jelenségnek tekinthető.

Mivel a papír-ceruza tesztek a populáció középtartományát mérik nagy pontossággal, ezért elképzelhető, hogy a nagyon alacsony és nagyon magas nehézségű itemekből kevés van a szélsőséges képességfejlettségű tanulók szimulációjának teljes lefuttatásához. Ebben az esetben értékelhetők a rövid szimulációk, főleg a *e*. kutatási kérdés tekintetében, illetve a kutatás szimulált bemeneti adatokkal kerülhet kiegészítésre.

A kutatás érvényességét befolyásolhatja, hogy az adatok papír-ceruza tesztekből származnak és számítógépes mérésre vonunk le következtetéseket. A médiahatás OKM-en történt vizsgálata nélkül (ld. 2.5 és 6.1 fejezetek) némiképp óvatosan kell megfogalmaznunk a következtetéseket, azonban a szimulációs technika általánosan használatos adaptív mérések előkészítésében és vizsgálatában.

A meredekebb itemekre alapozott szimuláció esetében realisztikusabb lett volna, ha az OKM-hez hasonlóan, háromparaméteres modell szerint generáltam volna az itembankot, illetve a korábbi szimulációk eredményeire támaszkodva Bayes-becslés helyett maximum likelihood, vagy az OKM-hez jobban illeszkedő expected a-posteriori becslést alkalmaztam volna.

A mérési területek és általában a disszertáció tekintetében korlát, hogy a számítógépes adaptív tesztelésre szorítkozik, holott a hazai és nemzetközi tapasztalatok azt mutatják, hogy a szövegértés tesztek (egy szöveghez tartozó itemek) felépítése miatt a többszakaszos adaptív tesztelés (MST) megfelelőbb. Ennek ellenére több érv is szól a CAT vizsgálata mellett. Egyrészt, mindkét módszer (CAT és MST) vizsgálata túlfeszítette volna a disszertáció kereteit, másrészt a többszakaszos adaptív tesztelésnek

már van hazai szakirodalma a Szegedi Tudományegyetem kutatásai keretében (ld. 3.4.2 fejezet), harmadrészt az ottani kutatások egyik kitekintése éppen a CAT-tal kapcsolatos vizsgálatokat nevezi meg továbblépési lehetőségként (Magyar, 2015), negyedrészt az OKM-nek egyre több mérési területe van, így lehetségesnek tartom, hogy bizonyos területek mérése többszakaszos, más területei számítógépes adaptív mérés keretében valósuljanak meg. Emiatt releváns lehet az OKM esetében és elsősorban a szövegértés területen az MST vizsgálata, a mérés céljához, a képességszintekhez, szövegtípusokhoz és gondolkodási műveletekhez illeszkedő szerkezet tervezése. Amennyiben a fejlesztések abba az irányba indulnak, hogy a teszt belépési értékének megválasztása alapuljon egyéb információkon (iskolai vagy demográfiai jellemzőkön, korábbi teljesítményen vagy más mérési terület eredményén), akkor a rugalmasabban indítható CAT alkalmasabb lehet rövidebb tesztek megvalósítására, mint a kötöttebb szerkezetű MST. Ugyanakkor, ha a mérési terület megköveteli, az MST is megvalósítható lehet alternatív nehézségű párhuzamos tesztekkel, hasonlóan a TIMSS és PIRLS csoportadaptív megvalósításához.

Eredményeim alapján az OKM esetében valószínűleg lehetséges a meglevő itembankra építve számítógépes, sőt adaptív mérésre áttérni. Thompson és Weiss (2011) számítógépes adaptív tesztek fejlesztésére vonatkozó keretrendszere szerint a fejlesztés következő lépései az itembank tartalmának felépítése vagy meglevő bank fejlesztése, majd az itemek bemérése. A jelenlegi itembank fejlesztésének szakaszai eredményeim alapján 1) a médiahatás pontos bemérése, valamint tesztelése alcsoportokon, 2) a nyílt itemek jellemzőinek összehasonlítása automatikus kódolású nyílt vagy zárt alternatívájuk jellemzőivel, 3) a nem automatizálható nyílt itemek elhagyásának empirikus vizsgálata, különösen a nagyon alacsony és nagyon magas képességfejlettség esetén, 4) a képességskála széleit mérő, nagyon könnyű és nagyon nehéz itemek fejlesztése, 5) a jelenlegi itemeknél nagyobb diszkrimináló képességgel rendelkező itemek fejlesztése. A fejlesztés ezen feladatai közül 1)–3) önálló, külső kutatásként is megvalósítható, míg 4) és 5) szakaszt a mérés biztonsága miatt annak szervezője, az Oktatási Hivatal tudja elvégezni.

A feladatfejlesztés egyik lépése a differenciált itemműködés vizsgálata, vagyis annak felfedése, hogy egy item valamely alcsoportok körében eltérő jellemzőkkel bír-e, ami azt jelentené, hogy adott körben az item alapján számított teljesítmény pontatlan. Mivel adaptív mérés esetén célzottan kerül kiosztásra az item, különösen fontos, hogy a paraméterek pontosak legyenek. Ugyanakkor az esetlegesen feltárt differenciált itemműködés nem feltétlenül hátrány. Amennyiben az alcsoport a mérés kezdete előtt

ismert, úgy lehetséges ugyanannak az itemnek különböző paraméterezésű változatait nyilvántartani, de az egyes változatok csak bizonyos alcsoportokban alkalmazhatók.

Az OKM 2023-tól már az 5–11. évfolyamokon kerül megszervezésre, ami felveti azt a kérdést, hogy az évfolyamok közös itemeire alkalmazhatók-e ugyanazok a paraméterek. Az évfolyamok és a mérési évek képességskáláinak összekötésére korábban a kiegészítő mérés állandó és jól működő itemei szolgáltak, ezek helyét a mérésben szereplő közös híd feladatok vették át. A mért évfolyamok kiterjesztése miatt érdemes lehet az itemek évfolyamonkénti differenciált itemműködését vizsgálni, esetleg az alcsoportoknál alkalmazható alternatív paraméterezést bevezetni az egyes évfolyamok vagy évfolyamok csoportjai esetében.

Jelenleg az OKM itemeit egydimenziósnak tekintjük, azaz minden item egy mérési területhez tartozik. A tesztfejlesztés során azokat az itemeket, melyek nem jól illeszkednek, mert a matematikai eszköztudás item szöveges komponense magasabb szövegértés képességfejlettséget igényel, vagy a szövegértés item nagyobb logikai erőfeszítést követel, jellemzően eltávolítják. A többdimenziós itemek fejlesztése azonban előrelépési lehetőség, cél is lehet abban az értelemben, hogy a később sorra kerülő mérési területek teljesítményszámításába korábbi területek itemeit is figyelembe lehetne venni. A többdimenziós IRT és a többdimenziós adaptív tesztek vizsgálata releváns, ugyanakkor friss kutatási irányok, a nemzetközi tanulóiteljesítmény-mérések esetében még nincsenek előjelei ezen irány mérlegelésének, így az OKM szempontjából is inkább távolabbinak mondható.

Irodalom

20/2012. (VIII. 31.) EMMI rendelet a nevelési-oktatási intézmények működéséről és a köznevelési intézmények névhasználatáról (2012). http://njt.hu/cgi_bin/njt_doc.cgi?docid=154155

30/2023. (VIII. 22.) BM rendelet - a 2023/2024. tanév rendjéről (2023). https://njt.hu/jogszabaly/2023-30-20-0A.0

120/2015. (V. 21.) Korm. rendelet A Klebelsberg Intézményfenntartó Központ fenntartásában működő egyes szakképzési feladatot ellátó köznevelési intézmények fenntartóváltásával összefüggő intézkedésekről, Pub. L. No. 120/2015., 69 Magyar közlöny 6080 (2015). http://www.kozlonyok.hu/nkonline/MKPDF/hiteles/mk15069.pdf

Akhtar, H., Silfiasari, Vekety, B., & Kovacs, K. (2023). The Effect of Computerized Adaptive Testing on Motivation and Anxiety: A Systematic Review and Meta-Analysis. *Assessment*, *30*(5), 1379–1390. https://doi.org/10.1177/10731911221100995

Anselmi, P., Colledani, D., & Robusto, E. (2019). A Comparison of Classical and Modern Measures of Internal Consistency. *Frontiers in Psychology*, *10*, 2714. https://doi.org/10.3389/fpsyg.2019.02714

Araci, F. G. İ., & Tan, Ş. (2022). Multidimensional Computerized Adaptive Testing Simulations in R. *International Journal of Assessment Tools in Education*, *9*(1), 118–137. https://doi.org/10.21449/ijate.909616

Arensman, R. M., Pisters, M. F., de Man-van Ginkel, J. M., Schuurmans, M. J., Jette, A. M., & de Bie, R. A. (2016). Translation, Validation, and Reliability of the Dutch Late-Life Function and Disability Instrument Computer Adaptive Test. *Physical Therapy*, *96*(9), 1430–1437. https://doi.org/10.2522/ptj.20150265

Australian Curriculum, Assessment and Reporting Authority. (2020). *National Assessment Program – Literacy and Numeracy 2019: Technical Report*. ACARA. https://nap.edu.au/docs/default-source/resources/naplan-2019_technical-report_final.pdf

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2014). *Tailored test design study 2013: Summary research report* (o. 18). ACARA. https://nap.edu.au/docs/default-source/default-document-library/tailored_test_design_study_2013_summary_research_report.pdf

Auxné Bánfi I., Balázsi I., Balkányi P., Balogh V. K., Gyapay J., Lak Á. R., Ostorics L. I., Palincsár I., Rábainé Szabó A., Rózsa C., Szabó L. D., Szepesi I., Szipőcsné Krolopp J., & Vadász C. (2014). *Országos kompetenciamérés—Technikai leírás*. Jelentés. Budapest: Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2012/OKM_Technikaileiras.pdf

Balázsi I. (2016). A hozzáadottérték-modellek alkalmazása a tanulói teljesítménymérésben. *Magyar Pedagógia, 116*(1), Article 1. https://doi.org/10.17670/MPed.2016.1.3

Balázsi, I., Balkányi, P., Balogh, V. K., Gyapay, J., Ostorics, L., Palincsár, I., Rábainé Szabó, A., Suhajda, E., Szepesi, I., Szipőcsné Krolopp, J., & Velkey, K. (2021). *Folytonosság és változás az Országos kompetenciamérés szövegértés és matematika tartalmi kereteiben* (Köt. 1). Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/digitalis_orszmer/OKMtartalmikeret_Szovegertes_Matematika.pdf

Balázsi I., Balkányi P., Ostorics L., Palincsár I., Rábainé Szabó A., Szepesi I., Szipőcsné Krolopp J., & Vadász C. (2014). *Az Országos kompetenciamérés tartalmi keretei—Szövegértés, matematika, háttérkérdőívek*. Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2014/AzOKMtartalmikeretei.pdf

Balázsi, I., Balkányi, P., & Vadász, C. (2017). *PIRLS 2016 Összefoglaló jelentés a 4. Évfolyamos tanulók eredményeiről* (0 kiad.). Oktatási Hivatal; MTMT. https://m2.mtmt.hu/api/publication/3340776

Balázsi I., Felvégi E., Rábainé Szabó A., & Szepesi I. (2009, július 17). *A 2006-os Országos kompetenciamérés tartalmi kerete* [Oktatáskutató és Fejlesztő Intézet]. https://ofi.hu/balazsi-ildiko-felvegi-emese-rabaine-szabo-annamaria-szepesi-ildiko-2006-os-orszagos

Balázsi, I., & Ostorics, L. (2011). *PISA2009 Digitális szövegértés. Olvasás a világhálón* (0 kiad.). Oktatási Hivatal; MTMT. https://m2.mtmt.hu/api/publication/2411815

Balázsi I., Ostorics L., Szalay B., Szepesi I., & Vadász C. (2013). *PISA 2012 Összefoglaló jelentés* (o. 80). Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/nemzetkozi_meresek/pisa/pisa2012_osszefoglalo_jelentes.pdf

Balázsi, I., Rábainé Szabó, A., Szabó, V., & Szepesi, I. (2005). A 2004-es Országos kompetenciamérés eredményei | Oktatáskutató és Fejlesztő Intézet. *Új Pedagógiai Szemle, 55*(12), 3–21.

Balázsi, I., & Zempléni, A. (2004). A hozottérték-index és a hozzáadott pedagógiai érték számítása a 2003-as kompetenciamérésben. *Új Pedagógiai Szemle, 54*(12), 36–50.

Balkányi, P., Gyapay, J., Lak, Á. R., Szabó Rábainé, A., Suhajda, E., Szabó, L. D., & Takácsné Kárász, J. (2018). *Országos kompetenciamérés 2017 Feladatok és jellemzőik szövegértés 6. Évfolyam*. Oktatási Hivatal, Köznevelési Mérés Értékelési Osztály. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2017/OKM2017_Feladatok_es_jellemzoik_Szovegertes_6.pdf

Balogh, V. K., Faddiné Buza, J., Nagyné Németh, B., Ostorics, L., & Szalay, B. (2021). *TERMÉSZETTUDOMÁNY az Országos kompetenciamérés új műveltségi területe* (Köt. 2). Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/digitalis_orszmer/OKMtartalmikeret_Termeszettudomany.pdf

Balogh, V. K., Garay-Madarász, Á., Havas, E., & Tankó, G. (2021). *Megújuló NYELVI MÉRÉSEK digitális felületen* (Köt. 3). Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/digitalis_orszmer/OKMtartalmikeret_Nyelvi.pdf

Bander K., Ercsei K., Galántai J., Gyökös E., Kurucz O., Nikitscher P., Szabó Z. A., & Szemerszki M. (2015). *Minőség és eredményesség a közoktatásban*. Oktatáskutató és Fejlesztő Intézet.

Barrada, J. R., Mazuela, P., & Olea, J. (2006). Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema, 18*(1), 156–159.

Belinszki Bálint, Szepesi Ildikó, Takácsné Kárász Judit, & Vadász Csaba. (2020). *Országos jelentés 2019* (Országos kompetenciamérés, o. 108). Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2019/Orszagos_jelentes_2019.pdf

Belov, D. (2014). Detecting Item Preknowledge in Computerized Adaptive Testing Using Information Theory and Combinatorial Optimization. *Journal of Computerized Adaptive Testing, 2*(3), 37–58. https://doi.org/10.7333/1410-0203037

Berliner, D. (2002). Educational research: The hardest science of them all. *Educational Researcher*, *31*, 18–20.

Biesta, G. (2009). Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability(Formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 33–46. https://doi.org/10.1007/s11092-008-9064-9

Biesta, G. J. J. (2011). *Good Education in an Age of Measurement: Ethics, Politics, Democracy*. Routledge.

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In F. M. Lord & M. R. Novick (Szerk.), *Statistical theories of mental test scores* (o. 397–479). Addison-Wesley.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, *6*(2), 258–276. https://doi.org/10.1016/0022-2496(69)90005-4

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Bock, R. D., & Zimowski, M. F. (2003). *Feasibility Studies of Two-Stage Testing in Large Scale Educational Assessment: Implications for NAEP* (Working Paper NCES2003-14; NAEP Validity Studies, o. 59). U.S. Department of Education, National Center for Education Statistics. https://eric.ed.gov/?id=ED478975

Brassil, C. E., & Couch, B. A. (2019). Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: A Bayesian item response model comparison. *International Journal of STEM Education*, *6*(16), 1–17. https://doi.org/10.1186/s40594-019-0169-0

Bridgeman, B. (1992). A Comparison of Quantitative Questions in Open-Ended and Multiple-Choice Formats. *Journal of Educational Measurement*, *29*(3), 253–271. https://doi.org/10.1111/j.1745-3984.1992.tb00377.x

Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation*, *62*, 1–9. https://doi.org/10.1016/j.stueduc.2019.04.005

Canz, T., Hoffmann, L., & Kania, R. (2020). Presentation-mode effects in large-scale writing assessments. *Assessing Writing*, *45*, 100470. https://doi.org/10.1016/j.asw.2020.100470

Chalhoub-Deville, M. (Szerk.). (1999). *Issues in computer-adaptive testing of reading proficiency*. Cambridge University Press.

Chang, H.-H. (2015). Psychometrics Behind Computerized Adaptive Testing. *Psychometrika*, *80*(1), 1–20. https://doi.org/10.1007/s11336-014-9401-5

Cho, E. (2016). Making Reliability Reliable: A Systematic Approach to Reliability Coefficients. *Organizational Research Methods*, *19*(4), 651–682. https://doi.org/10.1177/1094428116656239

Choe, E. M., & Fu, Y. (2018). Computerized Adaptive and Multistage Testing with R: Using Packages catR and mstR. *Measurement: Interdisciplinary Research and Perspectives*, *16*(4), 264–267. https://doi.org/10.1080/15366367.2018.1520560

Cizek, G. J., & Wollack, J. A. (Szerk.). (2016). *Handbook of Quantitative Methods for Detecting Cheating on Tests* (1. kiad.). Routledge. https://doi.org/10.4324/9781315743097

Colvin, K., Keller, L., & Robin, F. (2016). Effect of Imprecise Parameter Estimation on Ability Estimation in a Multistage Test in an Automatic Item Generation Context. *Journal of Computerized Adaptive Testing*, *4*(1), 1–18. https://doi.org/10.7333/1608-040101

Commission of the European Communities. (2010). *Towards more knowledge-based policy and practice in education and training Commission staff working document.* European Communities. https://publications.europa.eu/en/publication-detail/-/publication/962e3b89-c546-4680-ac84-777f8f10c590

Crins, M. H. P., van der Wees, P. J., Klausch, T., van Dulmen, S. A., Roorda, L. D., & Terwee, C. B. (2018). Psychometric properties of the PROMIS Physical Function item bank in patients receiving physical therapy. *PLoS ONE*, *13*(2), 1–14. https://doi.org/10.1371/journal.pone.0192187

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika*, *12*(1), 1–16. https://doi.org/10.1007/BF02289289

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), Article 3. https://doi.org/10.1007/BF02310555

Csányi R., & Molnár G. (2021). A tesztmegoldási motiváció kérdőíves és logadat alapú mérésének összehasonlító elemzése alacsony téttel rendelkező interaktív

problémamegoldó környezetben. *Magyar Pedagógia, 121*(3), 281–307. https://doi.org/10.17670/MPed.2021.3.281

Csapó B., Molnár G., & R. Tóth K. (2008). A papíralapú tesztektől a számítógépes adaptív tesztelésig. *Iskolakultúra*, *18*(3–4), 3–16.

Csíkos C., & Vidákovich T. (2012). A matematikatudás alakulása az empirikus vizsgálatok tükrében. In Csapó B. (Szerk.), *Mérlegen a magyar iskola* (o. 83–130). Nemzeti Tankönyvkiadó.

D. Molnár É., Molnár E., & Józsa K. (2012). Az olvasásvizsgálatok eredményei. In Csapó B. (Szerk.), *Mérlegen a magyar iskola* (o. 17–81). Nemzeti Tankönyvkiadó.

Daro, P., Stancavage, F., Ortega, M., DeStefano, L., & Linn, R. (2007). Validity Study of the NAEP Mathematics Assessment: Grades 4 and 8. In *American Institutes for Research*. American Institutes for Research. https://eric.ed.gov/?id=ED499213

Dennis, D. V. (2017). Learning From the Past: What ESSA Has the Chance to Get Right. *Reading Teacher*, *70*(4), 395–400. https://doi.org/10.1002/trtr.1538

Dobó, K. (2000). Az elektronikus szürke irodalom új formái, témái és felhasználása—2000. 46. Évf. 4. Szám—Elektronikus Periodika Archívum. *Könyvtári Figyelő*, *46*(4), 581–585.

DuToit, M. (Szerk.). (2003). *IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact*. Scientific Software International.

Educational Assessment Australia (EAA). (2013). *Tailored test design study 2013: Cognitive interviews* (o. 85). University os New South Wales. https://nap.edu.au/docs/default-source/default-document-library/nasop-tailored-test-design-study-2013-cognitive-interviews.pdf

ELTE. (2015). *„OKOS KÖZNEVELÉS" Javaslat a Nemzeti Oktatási Innovációs Rendszer stratégiájának kiegészítésére. NOIR+ stratégia.* ELTE PPK.

ETS. (2023, 0). *TOEFL Essentials Information Bulletin 2023-2024*. Https://Www.Ets.Org/Content/Dam/Ets-Org/Pdfs/Toefl/Toefl-Essentials-Bulletin.Pdf. https://www.ets.org/content/dam/ets-org/pdfs/toefl/toefl-essentials-bulletin.pdf

Every Student Succeeds Act of 2015, Pub. L. No. PUBLIC LAW 114–95—DEC. 10, 2015 (2015). https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Fehérvári, A., & Széll, K. (2014). Méltányosság az oktatásban: Tanulói eredmények, szülők, iskola. In K. Széll (Szerk.), *Az OECD az oktatásról—Adatok, elemzések,*

értelmezések (o. 41–52). Oktatáskutató és Fejlesztő Intézet. https://ofi.oh.gov.hu/sites/default/files/attachments/az_oecd_az_oktatasrol_ofi_2014.pdf

Finkelman, M. D., Kim, W., Weissman, A., & Cook, R. J. (2014). Cognitive Diagnostic Models and Computerized Adaptive Testing: Two New Item-Selection Methods That Incorporate Response Times. *Journal of Computerized Adaptive Testing*, *2*(4), 59–76. https://doi.org/10.7333/1412-0204059

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 Item Equivalence Study: Examining Mode Effects for Computer-Based Assessment and Implications for Measuring Trends. *Large-Scale Assessments in Education*, *6*. https://doi.org/10.1186/s40536-018-0064-z

Frey, A., & Seitz, N.-N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, *35*(2–3), 89–94. https://doi.org/10.1016/j.stueduc.2009.10.007

Frey, A., Seitz, N.-N., & Brandt, S. (2016). Testlet-Based Multidimensional Adaptive Testing. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01758

Geer, J., G. (1991). Do Open-ended questions measure „salient" issues? *Public Opinion Quarterly*, *55*(3), 360–370. https://doi.org/10.1086/269268

Gergely, B., & Takács, S. (2023). ATOM - Flexible multi-method machine learning framework to predict occupational success. *Alkalmazott Pszichológia*, *25*(3), 17–32. https://doi.org/10.17627/ALKPSZICH.2023.3.17

Gonthier, C., Aubry, A., & Bourdin, B. (2018). Measuring working memory capacity in children using adaptive tasks: Example validation of an adaptive complex span. *Behavior Research Methods*, *50*(3), 910–921. https://doi.org/10.3758/s13428-017-0916-4

Graham, A. K., Minc, A., Staab, E., Beiser, D. G., Gibbons, R. D., & Laiteerapong, N. (2019). Validation of the Computerized Adaptive Test for Mental Health in Primary Care. *The Annals of Family Medicine*, *17*(1), 23–30. https://doi.org/10.1370/afm.2316

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018

Halász G. (2013). *Az oktatáskutatás globális trendjei*. ELTE Eötvös Kiadó.

Hamhuis, E., Glas, C., & Meelissen, M. (2020). Tablet assessment in primary education: Are there performance differences between TIMSS' paper-and-pencil test and tablet test among Dutch grade-four students? *British Journal of Educational Technology*, *51*(6), 2340–2358. https://doi.org/10.1111/bjet.12914

Han, K. (Chris) T. (2020). Framework for Developing Multistage Testing With Intersectional Routing for Short-Length Tests. *Applied Psychological Measurement*, *44*(2), 87–102. https://doi.org/10.1177/0146621619837226

Harrison, P. M. C., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-03586-z

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2–3), 57–63. https://doi.org/10.1016/j.stueduc.2009.10.002

Hayes, A. F., & Coutts, J. J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But…. *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Herczegné Goldschmidt Z. (2016). Papíralapú és számítógép-alapú tesztelés összehasonlító vizsgálata olvasás-szövegértés területén 4. Évfolyamos diákok körében. *Iskolakultúra*, *26*(6), Article 6. https://doi.org/10.17543/ISKKULT.2016.6.30

Hicks, M. M. (1989). The Toefl Computerized Placement Test: Adaptive Conventional Measurement. *ETS Research Report Series*, *1989*(1), i–29. https://doi.org/10.1002/j.2330-8516.1989.tb00338.x

Hood, C. (1991). A Public Management for All Seasons? *Public Administration*, *69*(1), 3–19. https://doi.org/10.1111/j.1467-9299.1991.tb00779.x

Horn, D. (2010). *Elszámoltathatósági rendszerek elméleti háttere és nemzetközi tapasztalatai. Zárótanulmány*. Az MTA-KTI „A közoktatás teljesítményének mérése-értékelése, az iskolák elszámoltathatósága" programjának ACC 1503. számú produktuma. http://econ.core.hu/file/download//acc1503.doc

Hornke, L. F. (2000). Item Response Times in Computerized Adaptive Testing. *Psicológica*, *21*(1), 175–189.

Horváth, G. (1997). *A modern tesztmodellek alkalmazása* (S. Illyés, Szerk.). Akadémiai Kiadó.

Hülber L. (2012). A papír- és a számítógép alapú tesztelés összehasonlító vizsgálata különböző item paraméterek mentén. *Iskolakultúra*, *22*(12), Article 12.

Hülber L., & Molnár G. (2013). Papír és számítógép alapú tesztelés nagymintás összehasonlító vizsgálata matematika területén, 1-6. Évfolyamon. *Magyar Pedagógia*, *113*(4), 243–263.

Ito, K., & Segall, D. O. (2013). A Comparison of Four Methods for Obtaining Information Functions for Scores From Computerized Adaptive Tests With Normally Distributed Item Difficulties and Discriminations. *Journal of Computerized Adaptive Testing*, *1*(5).

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice*, *23*(4), 495–518. https://doi.org/10.1080/0969594X.2016.1147420

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education*, *44*(4), 476–493. https://doi.org/10.1080/03054985.2018.1430025

Jewsbury, P. A., & van Rijn, P. W. (2020). IRT and MIRT Models for Item Parameter Estimation With Multidimensional Multistage Tests. *Journal of Educational and Behavioral Statistics*, *45*(4), 383–402. https://doi.org/10.3102/1076998619881790

Karkó, Á. (2023). Mindig minden változik: Az emberek, a társadalom és az oktatás. *Új köznevelés*, *79*(7), 3–5.

Kehl D. (2012). Monte-Carlo-módszerek a statisztikában. *Statisztikai Szemle*, *90*(6), 521–543.

Kent, T. H., & Albanese, M. A. (1987). A Comparison of the Relative Efficiency and Validity of Tailored Tests and Conventional Quizzes. *Evaluation and the Health Professions*, *10*(1), 67–79. https://doi.org/10.1177/016327878701000106

Kertesi G. (2008). A közoktatási intézmények teljesítményének mérése-értékelése, az iskolák elszámoltathatósága. In Fazekas K., Köllő J., & Varga J. (Szerk.), *Zöld Könyv a magyar közoktatás megújításáért 2008* (o. 167–189). Ecostat Kormányzati Gazdasági- és Társadalom-statisztikai Kutató Intézet.

Kingsbury, G. G. (2009). Adaptive Item Calibration: A Process for Estimating Item Parameters Within a Computerized Adaptive Test. *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing, 1*(1), Article 1.

Kingsbury, G. G., & Hauser, C. (2004, április 13). *Computerized Adaptive Testing and No Child Left Behind* [Konferencia cikk]. 2004 Annual Meeting of the American Educational Research Association, San Diego, CA. https://www.nwea.org/uploads/2004/03/Computerized-Adaptive-Testing-and-NCLB.pdf

Kispál S., & Gergely B. (2022). Eltérő itemműködés vizsgálata az Országos kompetenciamérésben halmozottan hátrányos helyzetű diákok körében többdimenziós IRT modellek segítségével. *Psychologia Hungarica Caroliensis, 8*(4), 150–179. https://doi.org/10.52993/PSYHUNG.8.2020.4.4

Komatsu, H., & Rappleye, J. (2017). Did the shift to computer-based testing in PISA 2015 affect reading scores? A View from East Asia. *Compare: A Journal of Comparative and International Education, 47*(4), 616–623. https://doi.org/10.1080/03057925.2017.1309864

Kontra J. (2011). *A pedagógiai kutatások módszertana*. Kaposvári Egyetem. http://mek.niif.hu/12600/12648/12648.pdf

Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (2019). Construct Equivalence of PISA Reading Comprehension Measured With Paper-Based and Computer-Based Assessments. *Educational Measurement: Issues & Practice, 38*(3), 97–111. https://doi.org/10.1111/emip.12280

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160. https://doi.org/10.1007/BF02288391

Lak Á. R. (2020). A 2012. És 2015. Évi magyar PISA populációk összehasonlítása az Országos kompetenciamérés segítségével. *Köznevelési Elemzési Jelentések, 3*(1), 1–5.

Lak, Á. R., Palincsár, I., Szabó, L. D., Szepesi, I., Szipőcsné Krolopp, J., & Takácsné Kárász, J. (2018). *Országos kompetenciamérés 2017 Feladatok és jellemzőik matematika 6. Évfolyam*. Oktatási Hivatal, Köznevelési Mérés Értékelési Osztály. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2017/OKM2017_Feladatok_es_jellemzoik_Matematika_6.pdf

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lannert J. (2015). A PISA-adatok használata és értelmezése: A módszertani kritikák tükrében. *Educatio*, *24*(2), 18–28.

Lifelong Achievement Group, & Martin, A. J. (2015). *Online NAPLAN Testing and Student Motivation: Exploring Adaptive and Fixed Test Formats* (o. 45). Lifelong Achievement Group. https://www.nap.edu.au/_resources/Online_NAPLAN_and_Student_Motivation.PDF

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Inc.

Lu, J., & Wang, C. (2020). A Response Time Process Model for Not-Reached and Omitted Items. *Journal of Educational Measurement*, *57*(4), 584–620. https://doi.org/10.1111/jedm.12270

Luecht, R. M., & Nungester, R. J. (1998). Some Practical Examples of Computer-Adaptive Sequential Testing. *Journal of Educational Measurement*, *35*(3), 229–249. https://doi.org/10.1111/j.1745-3984.1998.tb00537.x

Magis, D., & Raîche, G. (2011). catR: An R Package for Computerized Adaptive Testing. *Applied Psychological Measurement*, *35*(7), 576–577. https://doi.org/10.1177/0146621611407482

Magis, D., & Raîche, G. (2012). Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software*, *48*(8), 1–31. https://doi.org/10.18637/jss.v048.i08

Magis, D., Yan, D., & von Davier, A. A. (2017a). An Overview of Computerized Adaptive Testing. In D. Magis, D. Yan, & A. A. von Davier (Szerk.), *Computerized Adaptive and Multistage Testing with R: Using Packages catR and mstR* (o. 35–51). Springer International Publishing. https://doi.org/10.1007/978-3-319-69218-0_3

Magis, D., Yan, D., & von Davier, A. A. (2017b). *Computerized Adaptive and Multistage Testing with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-69218-0

Magis, D., Yan, D., & von Davier, A. A. (2017c). Overview of Adaptive Testing. In D. Magis, D. Yan, & A. A. von Davier (Szerk.), *Computerized Adaptive and Multistage Testing with R: Using Packages catR and mstR* (o. 1–5). Springer International Publishing. https://doi.org/10.1007/978-3-319-69218-0_1

Magyar A. (2012). Számítógépes adaptív tesztelés. *Iskolakultúra*, *22*(6), 52–60.

Magyar A. (2014a). A szóolvasási készség adaptív mérését lehetővé tevő online tesztrendszer kidolgozása. *Magyar Pedagógia*, *114*(4), 259-279.

Magyar A. (2014b). Adaptív tesztek készítésének folyamata. *Iskolakultúra*, *14*(4), 26–35.

Magyar A. (2015). *Számítógép alapú adaptív és lineáris tesztek összehasonlító hatékonyságvizsgálata* [Doktori (PhD) értekezés, Szegedi Tudományegyetem]. https://doi.org/10.14232/phd.2633

Magyar A., & Molnár G. (2013). Számítógép alapú adaptív és rögzített formátumú tesztelés összehasonlító hatékonyságvizsgálat. *Magyar Pedagógia*, *113*(3), 181–193.

Magyar A., & Molnár G. (2015). A szóolvasási készség online mérésére kidolgozott adaptív és lineáris tesztrendszer összehasonlító hatékonyságvizsgálata. *Magyar Pedagógia*, *115*(4), Article 4. https://doi.org/10.17670/MPed.2015.4.403

Malkewitz, C. P., Schwall, P., Meesters, C., & Hardt, J. (2023). Estimating reliability: A comparison of Cronbach's α, McDonald's ωt and the greatest lower bound. *Social Sciences & Humanities Open*, *7*(1), 100368. https://doi.org/10.1016/j.ssaho.2022.100368

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Szerk.). (2016). *Methods and Procedures in TIMSS 2015* (https://timssandpirls.bc.edu/publications/timss/2015-methods.html). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/publications/timss/2015-methods.html

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Szerk.). (2017). *Methods and procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. https://timssandpirls.bc.edu/publications/pirls/2016-methods/P16_Methods_and_Procedures.pdf

Martin, M. O., Mullis, I. V. S., & von Davier, M. (2020). *Methods and Procedures: TIMSS 2019 Technical Report*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/pdf/TIMSS-2019-MP-Technical-Report.pdf

McDonald, R. P. (1999). *Test theory: A unified treatment* (o. xi, 485). Lawrence Erlbaum Associates Publishers.

Mead, A. D. (2006). An Introduction to Multistage Testing. *Applied Measurement in Education*, *19*(3), 185–187. https://doi.org/10.1207/s15324818ame1903_1

Molnár G. (2006). A Rasch-modell alkalmazása a társadalomtudományi kutatásokban. *Iskolakultúra*, *16*(12), 99–113.

Molnár G. (2010). Technológiaalapú mérésértékelés hazai és nemzetközi implementációi. *Iskolakultúra*, *20*(7–8), 22–34.

Molnár, G. (2015). A képességmérés dilemmái: A diagnosztikus mérések (eDia) szerepe és helye a magyar közoktatásban. *Géniusz M\Huhely: A magyar tehetségsegít\Ho szervezetek szövetsége (MATEHETSZ) kiadványsorozata*, *2*, 15–28.

Molnár G., & Csapó B. (2019). A diagnosztikus mérési rendszer technológiai keretei: Az eDia online platform. *Iskolakultúra*, *29*(4–5), 16–32. https://doi.org/10.14232/ISKKULT.2019.4-5.16

Molnár G., & Magyar A. (2015). A számítógép alapú tesztelés elfogadottsága pedagógusok és diákok körében. *Magyar Pedagógia*, *115*(1), 47–64. https://doi.org/10.17670/MPed.2015.1.47

Molnár, G., Magyar, A., Pásztor-Kovács, A., & Hülber, L. (2015). *A mérési-értékelési rendszer elektronikus alapokra helyezésével kapcsolatos helyzetelemzés* (o. 97). Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/unios_projektek/tamop318/OKM_kutatasi_eredmenyek2015/meresi_ertekelesi_rendszer.pdf

Mullis, I. V. S., & Martin, M. O. (Szerk.). (2015). *PIRLS 2016 Assessment Framework* (Second Edition). TIMSS & PIRLS, International Study Center, Lynch School of Education, Boston College. https://timssandpirls.bc.edu/pirls2016/downloads/P16_Framework_2ndEd.pdf

Mullis, I. V. S., & Martin, M. O. (Szerk.). (2017). *TIMSS 2019 Assessment Frameworks*. TIMSS & PIRLS. https://timss2019.org/wp-content/uploads/frameworks/T19-Assessment-Frameworks.pdf

Mullis, I. V. S., & Martin, M. O. (Szerk.). (2019). *PIRLS 2021 Assessment Frameworks*. TIMSS & PIRLS; Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/pirls2021/frameworks/

Mullis, I. V. S., Martin, M. O., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. https://timss2019.org/reports/wp-content/themes/timssandpirls/download-center/TIMSS-2019-International-Results-in-Mathematics-and-Science.pdf

Mullis, I. V. S., Martin, M. O., & von Davier, M. (Szerk.). (2021). *TIMSS 2023 Assessment Frameworks*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2023/frameworks/pdf/T23_Frameworks.pdf

Mullis, I. V. S., Von Davier, M., Foy, P., Fishbein, B., Reynolds, K., & Wry, E. (2023). *PIRLS 2021 International Results in Reading* (https://pirls2021.org/wp-content/uploads/2022/files/PIRLS-2021-International-Results-in-Reading.pdf). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5342

Muraki, E., & Bock, R. D. (1991). *PARSCALE* [Software]. Scientific Software International.

Nagybányai-Nagy O. (2006a). A pszichológiai tesztek reliabilitása. In Rózsa S., Nagybányai-Nagy O., & Oláh A. (Szerk.), *A pszichológiai mérés alapjai: Elmélet, módszer és gyakorlati alkalmazás* (o. 103–116). Bölcsész Konzorcium.

Nagybányai-Nagy O. (2006b). A pszichológiai tesztek validitása. In Rózsa S., Nagybányai-Nagy O., & Oláh A. (Szerk.), *A pszichológiai mérés alapjai: Elmélet, módszer és gyakorlati alkalmazás* (o. 117–124). Bölcsész Konzorcium.

Nahalka I. (2015). Tanulói teljesítménymérések alkalmazhatósága a neveléstudományban. In Széll K. (Szerk.), *Mit mér a műszer?* (o. 23–36). Oktatáskutató és Fejlesztő Intézet. https://ofi.oh.gov.hu/sites/default/files/attachments/mit_mr_a_mszer.pdf

Nahalka I. (2018). *Ellentmondások a pedagógiai mérés és értékelés elméleteiben*. https://doi.org/10.15773/EKE.HABIL.2018.008

Nahalka, I. (2023a). Oktatás és társadalom. In I. Falus & I. Szűcs (Szerk.), *A didaktika kézikönyve—Elméleti alapok a tanítás tanulásához* (o. 61–106). Akadémiai Kiadó. https://mersz.hu/dokumentum/m872d__1

Nahalka I. (2023b). *Tesztpedagógia – Áldás és átok* [Kézirat].

Nahalka, I., & Sipos, J. (2016). Az iskola eredményességével kapcsolatos nézetek. In Á. Vámos (Szerk.), *Tanuló pedagógusok és az iskola szakmai tőkéje* (o. 37–56). ELTE Eötvös Kiadó. https://www.eltereader.hu/media/2017/05/Vamos_Agnes_Tanulo_pedagogusok_READER.pdf

No Child Left Behind Act of 2001 (Passed Congress Version) (2002). https://www.govtrack.us/congress/bills/107/hr1/text

Nogami, Y., & Hayashi, N. (2010). A Japanese Adaptive Test of English as a Foreign Language: Developmental and Operational Aspects. In W. J. van der Linden & C. A. W. Glas (Szerk.), *Elements of Adaptive Testing* (o. 191–211). Springer New York. https://doi.org/10.1007/978-0-387-85461-8_10

OECD. (2009). *PISA 2009 Assessment Framework—Key Competencies in Reading, Mathematics and Science*. https://www.oecd.org/pisa/pisaproducts/44455820.pdf

OECD. (2013a). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD. https://doi.org/10.1787/9789264190511-en

OECD. (2013b). *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*. Organisation for Economic Co-operation and Development. https://doi.org/10.1787/9789264190658-en

OECD. (2014a). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)* (Köt. 1). OECD Publishing. https://doi.org/10.1787/9789264201118-en

OECD. (2014b). *PISA 2012 Technical Report* (https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf). OECD Publishing. https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

OECD. (2016a). Annex A6 The PISA 2015 field trial mode-effect study. In *PISA 2015 Results (Volume I): Excellence and Equity in Education*. OECD. https://doi.org/10.1787/9789264266490-en

OECD. (2016b). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. OECD. https://doi.org/10.1787/9789264266490-en

OECD. (2017a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving, revised Edition*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

OECD. (2017b). *PISA 2015 Technical Report* (https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf). OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf

OECD. (2019a). *PISA 2018 Assessment and Analytical Framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

OECD. (2019b). *PISA 2018 Results Combined Executive Summaries Volume I, II & III* (o. 354). PISA, OECD Publishing. https://www.oecd-ilibrary.org/docserver/5f07c754-

en.pdf?expires=1613894287&id=id&accname=guest&checksum=9633AA1B668DA2101BFBCFAD70053763

OECD. (2019c). *PISA 2018 Results (Volume I): What Students Know and Can Do*. OECD Publishing. https://doi.org/10.1787/5f07c754-en

OECD. (2019d). PISA 2018 Technical Report—Chapter 2 Test Design and Test Development. In *PISA 2018 Assessment and Analytical Framework*. https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD/DeSeCo. (2003). *Summary of the final report—"Key Competencies for a Successful Life and a Well-Functioning Society"*.

Oktatási Hivatal. (é. n.). *Változások az Országos kompetenciamérés skáláiban*. oktatas.hu. Elérés 2021. március 15., forrás https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2010/valt_orszmer_skala_110228.pdf

Oktatási Hivatal. (2019). *PISA 2018 Összefoglaló jelentés* (o. 99). Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/nemzetkozi_meresek/pisa/PISA2018_v6.pdf

Oktatási Hivatal. (2020). *OKM 2019 FIT-jelentés: Útmutató a Telephelyi jelentés ábráinak értelmezéséhez*. https://okm.kir.hu/fit/files/OKM2019_Utmutato_a_Telephelyi_jelentes_abrainak_ertelmezesehez.pdf

Oktatási Hivatal. (2021, szeptember 15). *A digitális országos mérések általános leírása*. https://www.oktatas.hu/kozneveles/meresek/digitalis_orszagos_meresek/altalanos_leiras

Oktatási Hivatal. (2022a). *OKM 2021 FIT-jelentés: Útmutató a Telephelyi jelentés ábráinak értelmezéséhez*. Oktatási Hivatal. https://okm.kir.hu/fit/files/OKM2021_Utmutato_a_Telephelyi_jelentes_abrainak_ertelmezesehez.pdf

Oktatási Hivatal. (2022b, augusztus 17). *A digitális országos mérések általános leírása*. https://www.oktatas.hu/kozneveles/meresek/digitalis_orszagos_meresek/altalanos_leiras

Oktatási Hivatal. (2023a). *Országos kompetenciamérés—Digitális országos mérések, Országos jelentés 2022* (Országos kompetenciamérés, o. 30) [Országos jelentés]. Oktatási Hivatal. https://okm.kir.hu/fit2/pdf/OKM_2022_Orszagos_jelentes.pdf

Oktatási Hivatal. (2023b). *Tájékoztató a 2023/2024. Tanévi digitális országos kompetenciamérésről*. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/digitalis_orszmer/Orszagos_kompetenciameres_2023_2024.pdf

Ostorics L., Szalay B., Szepesi I., & Vadász C. (2016). *PISA 2015 Összefoglaló jelentés* (o. 90). Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/nemzetkozi_meresek/pisa/PISA2015_osszefoglalo_jelentes.pdf

Ozturk, N. B., & Dogan, N. (2015). Investigating Item Exposure Control Methods in Computerized Adaptive Testing. *Educational Sciences: Theory & Practice*, *15*(1), 85–98. https://doi.org/10.12738/estp.2015.1.2593

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., & Moher, D. (2021). Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology*, *134*, 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003

Palincsár I., Szalay B., Szepesi I., Ostorics L., & Vadász C. (2020). *TIMSS 2019 Összefoglaló jelentés* (o. 184). Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/nemzetkozi_meresek/timss/TIMSS2019.pdf

Pásztor-Kovács A., Magyar A., Hülber L., Pásztor A., & Tongori Á. (2013). Áttérés online tesztelésre – a mérés-értékelés új dimenziói. *Iskolakultúra*, *23*(11), 86–100.

Peters, G.-J. Y. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality | European Health Psychologist. *The European Health Psychologist*, *16*(2), 56–69.

Petersen, M. Aa., Giesinger, J. M., Holzner, B., Arraras, J. I., Conroy, T., Gamper, E.-M., King, M. T., Verdonck-de Leeuw, I. M., Young, T., & Groenvold, M. (2013). Psychometric evaluation of the EORTC computerized adaptive test (CAT) fatigue item pool. *Quality of Life Research*, *22*(9), 2443–2454. https://doi.org/10.1007/s11136-013-0372-2

R Core Team. (2016). *R: A Language and Environment for Statistical Computing* [Software]. R Foundation for Statistical Computing. https://www.R-project.org/

R. Tóth, K., & Hódi, Á. (2011). Számítógépes és papír-ceruza teszteredmények összehasonlító vizsgálata az olvasás-szövegértés területén. *Magyar Pedagógia*, *111*(4), 313–332.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Robitzsch, A., Luedtke, O., Goldhammer, F., Kroehne, U., & Koeller, O. (2020). Reanalysis of the German PISA Data: A Comparison of Different Approaches for Trend Estimation With a Particular Emphasis on Mode Effects. In *Frontiers in Psychology* (Köt. 11). FRONTIERS MEDIA SA. https://doi.org/10.3389/fpsyg.2020.00884

Robitzsch, A., & Lüdtke, O. (2019). Linking errors in international large-scale assessments: Calculation of standard errors for trend estimation. *Assessment in Education: Principles, Policy & Practice*, *26*(4), Article 4. https://doi.org/10.1080/0969594X.2018.1433633

Rother, E. T. (2007). Systematic literature review X narrative review. *Acta Paulista de Enfermagem*, *20*, v–vi. https://doi.org/10.1590/S0103-21002007000200001

RStudio Team. (2020). *RStudio: Integrated Development for R* [Software]. RStudio, PBC. http://www.rstudio.com/

Rutkowski, L., & Valdivia, M. (2020). Review of *Computerized Adaptive and Multistage Testing With R: Using Packages* catR *and* mstR. *Journal of Educational and Behavioral Statistics*, *45*(1), 108–115. https://doi.org/10.3102/1076998619858626

Şahin, A., & Weiss, D. J. (2015). Effects of Calibration Sample Size and Item Bank Size on Ability Estimation in Computerized Adaptive Testing. *Educational Sciences: Theory & Practice*, *15*(6), 1585–1595. https://doi.org/10.12738/estp.2015.6.0102

Sari, H. İ. (2020). Testing Multistage Testing Configurations: Post-Hoc vs. Hybrid Simulations. *International Journal of Psychology and Educational Studies*, *7*(1), 27–37. https://doi.org/10.17220/ijpes.2020.01.003

Sari, H. İ., & Huggins-Manley, A. C. (2017). Examining Content Control in Adaptive Tests: Computerized Adaptive Testing vs. Computerized Adaptive Multistage Testing. *Educational Sciences: Theory & Practice*, *17*(5). https://doi.org/10.12738/estp.2017.5.0484

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Sciences Reform Act of 2002. To Provide for Improvement of Federal Education Research, Statistics, Evaluation, Information, and Dissemination, and for Other Purposes. (2002). https://www.govtrack.us/congress/bills/107/hr3801/text

Sie, H., Finkelman, M. D., Riley, B., & Smits, N. (2015). Utilizing Response Times in Computerized Classification Testing. *Applied Psychological Measurement*, *39*(5), 389–405. https://doi.org/10.1177/0146621615569504

Spineti, J. P., & Hambleton, R. K. (1977). A Computer Simulation Study of Tailored Testing Strategies for Objective-Based Instructional Programs. *Educational and Psychological Measurement*, *37*(1), 139–158. https://doi.org/10.1177/001316447703700115

Stafford, R. E., Runyon, C. R., Casabianca, J. M., & Dodd, B. G. (2019). Comparing computer adaptive testing stopping rules under the generalized partial-credit model. *Behavior Research Methods*, *51*(3), 1305–1320. https://doi.org/10.3758/s13428-018-1068-x

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Sympson, J. B., Weiss, D. J., & Ree, M. J. (1982). *Predictive Validity of Conventional and Adaptive Tests in an Air Force Training Environment.* MINNESOTA UNIV MINNEAPOLIS DEPT OF PSYCHOLOGY. https://apps.dtic.mil/docs/citations/ADA119031

Széll K. (2018). *Iskolai légkör és eredményesség: Fókuszban a reziliens és a veszélyeztetett iskolák*. Belvedere Meridionale. http://real-eod.mtak.hu/9458/

Szemerszki, M. (2014). Mérés és értékelés az oktatásban. In K. Széll (Szerk.), *Az OECD az oktatásról—Adatok, elemzések, értelmezések* (o. 17–28). Oktatáskutató és Fejlesztő Intézet. https://ofi.oh.gov.hu/sites/default/files/attachments/az_oecd_az_oktatasrol_ofi_2014.pdf

Szemerszki M. (2015). A tanulói teljesítménymérések szerepe a tényekre alapozott oktatáspolitikában. In Széll K. (Szerk.), *Mit mér a műszer?* (o. 9–22). Oktatáskutató és Fejlesztő Intézet. https://ofi.oh.gov.hu/sites/default/files/attachments/mit_mr_a_mszer.pdf

T. Kárász, J. (é. n.). *A számítógépes adaptív mérés módszertani fejlesztését segítő catR programcsomag főbb funkcióinak ismertetése* [Kézirat].

T. Kárász J., Nagybányai Nagy O., Széll K., & Takács S. (2022). Cronbach-alfa: Vele vagy nélküle? *Magyar Pszichológiai Szemle, 77*(1), 81–98. https://doi.org/10.1556/0016.2022.00004

T. Kárász J., & Széll K. (2023). Hogyan térnek el a papír-ceruza és számítógépes teszteredmények? - Szisztematikus szakirodalom áttekintés a PISA, TIMSS és PIRLS mérésekkel kapcsolatos tapasztalatokról. *Iskolakultúra, 33*(3), Article 3.

T. Kárász, J., Széll, K., & Takács, S. (2023). Closed formula of test length required for adaptive testing with medium probability of solution. *Quality Assurance in Education, 31*(4), 637–651. https://doi.org/10.1108/QAE-03-2023-0042

T. Kárász J., & Takács S. (2021). Adaptív tesztek minimális hosszának, hibájának, értékelési szintjének és a megoldók számának összefüggései—Általános megoldási aránnyal. *Alkalmazott Matematikai Lapok, 38*(1), 39–58. https://doi.org/10.37070/AML.2021.38.1.04

T. Kárász, J., & Takács, S. (2023). Use of open and closed items in automation of evaluation systems. *Alkalmazott Pszichológia, 25*(3), 33–54. https://doi.org/10.17627/ALKPSZICH.2023.3.33

The National Assessment Governing Board. (2017). *Mathematics Framework for the 2017 National Assessment of Educational Progress*. U.S. department of Education. https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/mathematics/2017-math-framework.pdf

Thompson, G. (2017). Computer adaptive testing, big data and algorithmic approaches to education. *British Journal of Sociology of Education, 38*(6), 827–840. https://doi.org/10.1080/01425692.2016.1158640

Thompson, N. A., & Weiss, D. A. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research, and Evaluation, 16*(1), 1–9. https://doi.org/10.7275/WQZT-9427

Tomasz, G. (2011). Tesztek helyett oktatást! - Avagy egy lakásfelújítás szellemi hozadéka. *Educatio, 20*(1), 123–127.

Tóth, E. (2010). Tesztalapú elszámoltathatóság a közoktatásban. *Iskolakultúra, 20*(1), 60–78.

Tóth E. (2014). *Pedagógusok véleménye a rendszerszintű mérésekről és azok tanítási folyamatra gyakorolt hatásáról* [Doktori (PhD) értekezés]. Szegedi Tudományegyetem.

Urry, V. W. (1970). *A Monte Carlo investigation of logistic test models* [Nem publikált PhD-értekezés, Purdue University]. https://files.eric.ed.gov/fulltext/ED058317.pdf

van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, *13*(1), Article 1. https://doi.org/10.1207/s15324818ame1301_2

van der Linden, W. J., & Ren, H. (2019). A Fast and Simple Algorithm for Bayesian Adaptive Testing. *Journal of Educational and Behavioral Statistics*, *45*(1), 58–85. https://doi.org/10.3102/1076998619858970

Vargha, A. (2015). *Matematikai statisztika*. Pólya Kiadó.

Vass, V. (2006). A kompetencia fogalmának értelmezése. In K. Demeter (Szerk.), *A kompetencia. Kihívások és értelmezések* (o. 139–161). Országos Közoktatási Intézet (OKI). https://ofi.oh.gov.hu/tudastar/hazai-fejlesztesi/kompetencia-fogalmanak

Velkey, K. (2018). A 2015-ös lengyel PISA-eredmények és ami mögöttük van. *Educatio*, *27*(2), 332–340. https://doi.org/10.1556/2063.27.2018.2.14

Volacu, A. (2018). Justice, Efficiency, and the New Public Management. *Australian Journal of Public Administration*, *77*(3), 404–414. https://doi.org/10.1111/1467-8500.12263

von Davier, M., Foy, P., Martin, M. O., & Mullis, I. V. S. (2020). Examining eTIMSS country differences between eTIMSS data and bridge Data: A look at country-level mode of administration effects. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Szerk.), *Methods and Procedures: TIMSS 2019 Technical Report* (o. 13.1-13.24). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/pdf/T19_MP_Ch13-etimss-country-differences.pdf#pagemode=none&page=1

von Davier, M., Mullis, I. V. S., Fishbein, B., & Foy, P. (Szerk.). (2023). *Methods and Procedures: PIRLS 2021 Technical Report*. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5892

von Helmholtz, H. (1977). Numbering and Measuring from an Epistemological Viewpoint. In *Epistemological Writings* (o. 72–114). Springer Netherlands. https://doi.org/10.1007/978-94-010-1115-0_3

Wang, X., Liu, Y., Robin, F., & Guo, H. (2019). A Comparison of Methods for Detecting Examinee Preknowledge of Items. *International Journal of Testing*, *19*(3), 207–226. https://doi.org/10.1080/15305058.2019.1610886

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Weiss, D. J. (2011). Better Data From Better Measurements Using Computerized Adaptive Testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1–27. https://doi.org/10.2458/v2i1.12351

Weiss, D. J., & Kingsbury, G. G. (1984). Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wise, S. L. (2014). The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees. *Journal of Computerized Adaptive Testing*, *2*(1), 1–17.

Wright, B. D. (1977). Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement*, *14*(2), Article 2.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Mesa Press.

Wright, B. D., & Stone, M. H. (1999). *Measurement Essentials 2nd Ed.* WIDE RANGE, INC.

Yamamoto, K., Khorramdel, L., & Shin, H. J. (2018). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling*, *60*(3), 347–368.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage Adaptive Testing Design in International Large-Scale Assessments. *Educational Measurement: Issues and Practice*, *37*(4), 16–27. https://doi.org/10.1111/emip.12226

Yang, L., & Reckase, M. D. (2020). The Optimal Item Pool Design in Multistage Computerized Adaptive Tests With the p-Optimality Method. *Educational and Psychological Measurement*, *80*(5), 955–974. https://doi.org/10.1177/0013164419901292

Zehner, F., Goldhammer, F., Lubaway, E., & Sälzer, C. (2019). Unattended consequences: How text responses alter alongside PISA's mode change from 2012 to 2015. *Education Inquiry*, *10*(1), 34–55. https://doi.org/10.1080/20004508.2018.1518080

Zehner, F., Kroehne, U., Hahnel, C., & Goldhammer, F. (2020). PISA reading: Mode effects unveiled in short text responses. *Psychological Test and Assessment*

Modeling, *62*(1), 85–105. https://www.proquest.com/scholarly-journals/pisa-reading-mode-effects-unveiled-short-text/docview/2387625745/se-2.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage Testing: Issues, Designs, and Research. In W. J. van der Linden & C. A. W. Glas (Szerk.), *Elements of Adaptive Testing* (o. 355–372). Springer New York. https://doi.org/10.1007/978-0-387-85461-8_18

Mellékletek

1. melléklet

A meredekebb itemekből álló feladatbankon futtatott szimuláció outputja

az egyes fázisok beállítása, az alapbeállításoktól eltérő elemek jelölve

```
> start0 <- list(nrItems = 1, theta = 1500, randomesque = 5, startSelect
   = "MFI" )
> test0 <- list(method = "BM", priorDist = "norm", priorPar = c(1500,
200), range = c(700, 2300), itemSelect = "MFI", randomesque = 5 )
> final0 <- list(method = "EAP", priorDist = "norm", priorPar = c(1500,
200), range = c(700, 2300), parInt = c(700, 2300, 33))
```

megállítási kritérium, 50 item vagy a becslés hibája 60 pont alá csökken

```
> stop1 <- list(rule = c("length","precision"), thr = c(50, 60))
```

adaptív teszt szimulációja a képességpontok vektorára és a válaszmintázat mátrixára

```
> simres0 <- simulateRespondents(thetas = theta0, itemBank = itPar0,
responsesMatrix = res0, model = NULL, cbControl = NULL, rmax = 0.2,
Mrmax = "restricted",
start = start0, test = test0, stop = stop1, final = final0,
save.output = TRUE, output = c("C:/catR_examples/", "simres0", "txt"))
> print(simres0)
 ** Post-hoc simulation of multiple examinees **

Simulation time: 1.9573 hours

Number of simulees: 90000
Item bank size: 300 items
IRT model: Two-Parameter Logistic model

Item selection criterion: MFI
 Stopping rules:
      Stopping criterion 1: length of test
         Maximum test length: 50
      Stopping criterion 2: precision of ability estimate
         Maximum SE value: 60
 rmax: 0.2
       Restriction method: restricted

Mean test length: 19.68907 items
Correlation(assigned thetas,CAT estimated thetas): 0.9535
RMSE: 60.4422
Bias: -0.3861
Maximum exposure rate: 0.2
Number of item(s) with maximum exposure rate: 44
Minimum exposure rate: 0
Number of item(s) with minimum exposure rate: 65
Item overlap rate: 0.174

Conditional results
                     Measure       D1       D2       D3       D4       D5
                  Mean Theta 1147.414 1289.263 1363.447 1422.483 1474.793
                        RMSE   66.663   59.834   59.054   57.958   57.599
                   Mean bias   32.046    18.71   12.686    6.703    2.109
              Mean test length    16.36   17.751   19.003   19.971   20.742
          Mean standard error   60.369   60.349   60.364   60.359   60.333
```

Proportion stop rule satisfied	1	1	1	1	1
Number of simulees	9000	9000	9000	9000	9000

Measure	D6	D7	D8	D9	D10
Mean Theta	1525.28	1576.87	1634.585	1708.696	1851.716
RMSE	57.769	57.986	59.205	61.043	66.432
Mean bias	-1.846	-9.074	-12.931	-19.701	-32.563
Mean test length	21.178	21.296	21.129	20.45	19.011
Mean standard error	60.337	60.319	60.301	60.292	60.278
Proportion stop rule satisfied	1	1	1	1	1
Number of simulees	9000	9000	9000	9000	9000

Megjegyzés. A jobb átláthatóság érdekében a *Conditional results* részben az eredeti outputot szerkezetileg módosítottam, a sor elején megismételve a fejlécet.

EÖTVÖS LORÁND TUDOMÁNYEGYETEM

ADATLAP a doktori értekezés nyilvánosságra hozatalához

I. A doktori értekezés adatai

A szerző neve: **Takácsné Kárász Judit**

A doktori értekezés címe és alcíme: **Adaptív teljesítménymérési algoritmusok kidolgozása az Országos kompetenciamérés adatainak felhasználásával**

A doktori iskola neve: **Neveléstudományi Doktori Iskola**

A doktori iskolán belüli doktori program neve: **Oktatás-tanulás-egyenlőtlenségek program**

A témavezető neve és tudományos fokozata: **Dr. habil. Nahalka István CSc, ny. egyetemi docens; Dr. habil. Széll Krisztián László, egyetemi docens**

A témavezető munkahelye: **-; ELTE PPK**

MTA Adatbázis-azonosító: **10067913**

DOI-azonosító[50]: **10.15476/ELTE.2024.139**

II. Nyilatkozatok

1. A doktori értekezés szerzőjeként[51]

a) hozzájárulok, hogy a doktori fokozat megszerzését követően a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az ELTE Digitális Intézményi Tudástárban. Felhatalmazom a **Neveléstudományi Doktori Iskola** hivatalának ügyintézőjét ……., hogy az értekezést és a téziseket feltöltse az ELTE Digitális Intézményi Tudástárba, és ennek során kitöltse a feltöltéshez szükséges nyilatkozatokat.

b) kérem, hogy a mellékelt kérelemben részletezett szabadalmi, illetőleg oltalmi bejelentés közzétételéig a doktori értekezést ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;[52]

c) kérem, hogy a nemzetbiztonsági okból minősített adatot tartalmazó doktori értekezést a minősítés (…….*dátum*)-ig tartó időtartama alatt ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;[53]

[50] A kari hivatal ügyintézője tölti ki.
[51] A megfelelő szöveg aláhúzandó.
[52] A doktori értekezés benyújtásával egyidejűleg be kell adni a tudományági doktori tanácshoz a szabadalmi, illetőleg oltalmi bejelentést tanúsító okiratot és a nyilvánosságra hozatal elhalasztása iránti kérelmet.
[53] A doktori értekezés benyújtásával egyidejűleg be kell nyújtani a minősített adatra vonatkozó közokiratot.

d) kérem, hogy a mű kiadására vonatkozó mellékelt kiadó szerződésre tekintettel a doktori értekezést a könyv megjelenéséig ne bocsássák nyilvánosságra az Egyetemi Könyvtárban, és az ELTE Digitális Intézményi Tudástárban csak a könyv bibliográfiai adatait tegyék közzé. Ha a könyv a fokozatszerzést követőn egy évig nem jelenik meg, hozzájárulok, hogy a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban.[54]

2. A doktori értekezés szerzőjeként kijelentem, hogy

a) a ELTE Digitális Intézményi Tudástárba feltöltendő doktori értekezés és a tézisek saját eredeti, önálló szellemi munkám és legjobb tudomásom szerint nem sértem vele senki szerzői jogait;

b) a doktori értekezés és a tézisek nyomtatott változatai és az elektronikus adathordozón benyújtott tartalmak (szöveg és ábrák) mindenben megegyeznek.

3. A doktori értekezés szerzőjeként hozzájárulok a doktori értekezés és a tézisek szövegének plágiumkereső adatbázisba helyezéséhez és plágiumellenőrző vizsgálatok lefuttatásához.

Kelt: Budapest, 2023.06.17.

a doktori értekezés szerzőjének aláírása

[54] A doktori értekezés benyújtásával egyidejűleg be kell nyújtani a mű kiadásáról szóló kiadói szerződést.